Histopathological diagnosis of breast cancer using machine learning

Babak Ehteshami Bejnordi

This book was typeset by the author using LaTeX2$_\varepsilon$.

Printed by IPSKAMP printing, Nijmegen.

Histopathological diagnosis of breast cancer using machine learning

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op
woensdag, 20 december 2017
om 10:30 uur precies

door

Babak Ehteshami Bejnordi

geboren op 18 September 1986
te Rasht, Guilan, Iran

Promotoren:	**Prof. dr. N. Karssemeijer**
Copromotor:	**Dr. J.A.W.M van der Laak** **Dr. G. Litjens**
Manuscriptcommissie:	**Prof. dr. T. Heskes** **Prof. dr. N. Rajpoot** (The University of Warwick, United Kingdom) **Prof. dr. J. Wesseling**

The research described in this thesis was carried out at the Diagnostic Image Analysis Group, Radboud University Medical Center (Nijmegen, the Netherlands).
This work was funded by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement n°601040 and is part of the VPH-PRISM project, Virtual Physiological Human: Personalized Predictive Breast Cancer Therapy Through Integrated Tissue Micro-Structure Modeling.

Financial support for publication of this thesis was kindly provided by Radboud University Medical Center.

Histopathological diagnosis of breast cancer using machine learning

Doctoral Thesis

To obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the rector magnificus prof. dr. J.H.J.M. van Krieken,
according to the decision of the Council of Deans
to be defended in public on
Wednesday, December 20, 2017
at 10:30am

by

Babak Ehteshami Bejnordi

born on September 18, 1986
in Rasht, Guilan, Iran

Supervisors:	**Prof. dr. N. Karssemeijer**
Co-supervisors:	**Dr. J.A.W.M van der Laak** **Dr. G. Litjens**
Manuscript committee:	**Prof. dr. T. Heskes** **Prof. dr. N. Rajpoot** (The University of Warwick, United Kingdom) **Prof. dr. J. Wesseling**

The research described in this thesis was carried out at the Diagnostic Image Analysis Group, Radboud University Medical Center (Nijmegen, the Netherlands).
This work was funded by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement n°601040 and is part of the VPH-PRISM project, Virtual Physiological Human: Personalized Predictive Breast Cancer Therapy Through Integrated Tissue Micro-Structure Modeling.

Financial support for publication of this thesis was kindly provided by Radboud University Medical Center.

TABLE OF CONTENTS

Introduction

1

The known is finite, the unknown infinite; intellectually we stand on an islet in the midst of an illimitable ocean of inexplicability. Our business in every generation is to reclaim a little more land, to add something to the extent and the solidity of our possessions

T. H. Huxley, On the Reception of the Origin of Species, 1887

1.1 Breast Cancer

Breast cancer is the second cause of cancer death among women worldwide, and is the leading cause of cancer death (death from any type of cancer) among women ages 20-39[1]. The natural history of breast cancer involves the progression through clinical and pathological stages starting with normal epithelial proliferation to invasive carcinoma via hyperplasia and carcinoma in-situ, and culmination in metastatic disease[2]. Figure 1.1 shows examples of various non-malignant and malignant lesions of the breast. Although the initiating steps and the precise underlying mechanism for tumorigenesis have not been fully elucidated, it appears that nearly all invasive breast cancers arise from in-situ carcinomas[2]. Multiple pathological and biological features distinguish in-situ carcinoma, invasive carcinoma, and benign proliferative breast lesions from each other. Accurate diagnosis of breast proliferative disease conditions is pivotal to determine the optimal treatment plan.

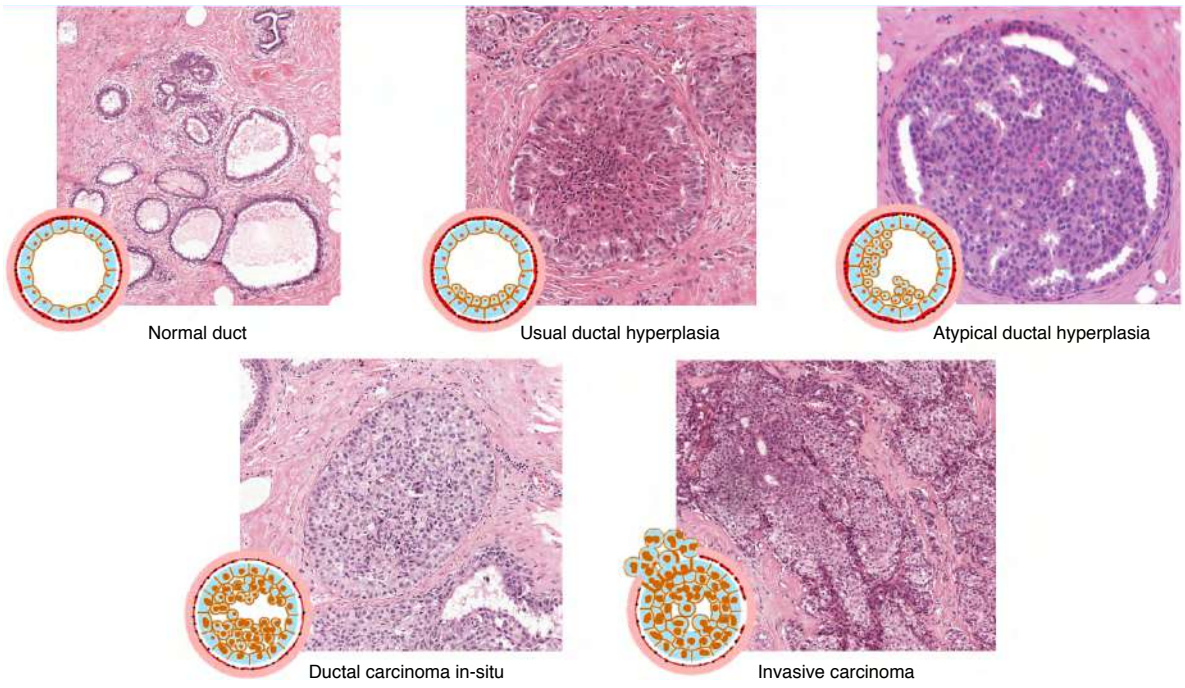

Figure 1.1: Progression of breast cancer at different stages. Breast cancer is thought to progress from atypical hyperplasia and ductal carcinoma in-situ (DCIS) to invasive cancer.

Currently, the diagnosis of these conditions largely depends on a careful examination of breast tissue sections under a microscope by a pathologist. The procedure for preparation of tissue sections consists of the following steps: (i) collecting the tissue removed from the body for diagnosis. (ii) applying a fixative to preserve the tissue. (iii) embedding the tissue in paraffin. (iv) cutting the tissue into thin (3-5 µm)

sections and mounting it on glass slides, and finally (v) treating the tissue with multiple contrasting stains to highlight different tissue structures and cellular features. These steps are usually performed by a pathology lab technician who then sends the specimen to a pathologist for interpretation.

Staining is one of the most important steps in tissue preparation. This is because most cells are colorless and transparent. Therefore, histological sections have to be stained to highlight important features of the tissue as well as to enhance the tissue contrast. Microscopical evaluation of such stained tissue sections provides invaluable information to the pathologists for diagnosing and characterizing various pathological conditions. Hematoxylin and eosin (H&E) staining is the most commonly used staining technique in histopathology. Hematoxylin stains cell nuclei blue, and the counter-stain eosin stains cytoplasmic and non-nuclear components in different shades of pink. H&E staining is non-specific, meaning that it stains most of the cells in much the same way. Specific staining techniques are those that selectively stain particular chemical groupings or molecules within cells or tissues. Immunohistochemistry (IHC), as an example, makes it possible to visualize the distribution and localization of specific cellular components within a cell or tissue. Figure 1.2 shows examples of H&E and IHC staining of a metastatic region in a lymph node tissue section. Clearly, the use of IHC in the right panel supports recognition of the tumor area.

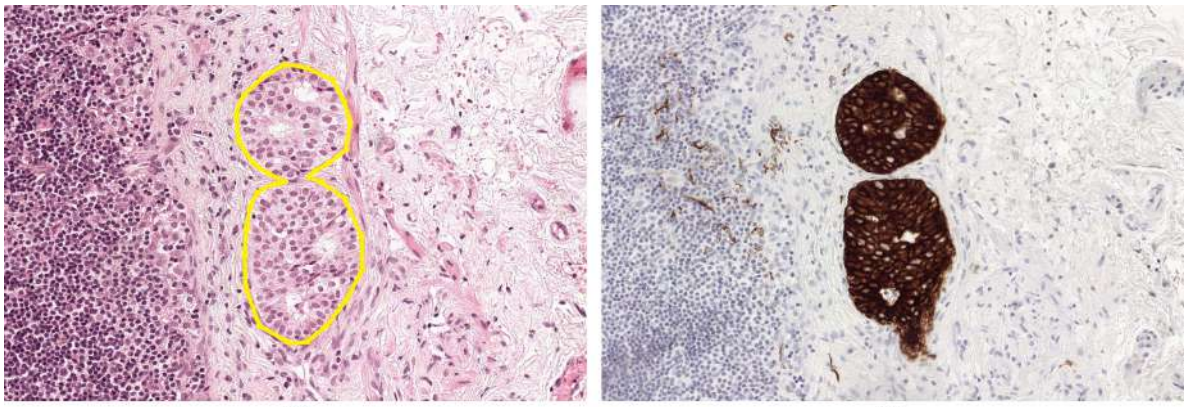

Figure 1.2: (a) Example of hematoxylin and eosin (H&E) staining of a metastatic region in a lymph node tissue and (b) the corresponding region detected by immunohistochemistry.

Histopathological analysis of breast tissue specimens almost always starts with a careful examination of H&E stained specimens. Despite the introduction of powerful and highly specific staining procedures in the last 40 years, H&E remains the most widely used tissue stain for determining the diagnosis and prognosis of breast can-

cer patients. The pathological classification of breast proliferative diseases is based on nuclear morphometric features and architectural patterns of cells and glands. Invasive tumors are routinely graded based on an assessment of tubule/gland formation, nuclear pleomorphism and mitotic count. Besides being laborious and time-consuming, classification and grading remain a challenge owing to differing pathological criteria, interobserver variability[3], and the heterogeneous tumor growth patterns. Whilst computerized analysis could potentially address the issue of inappropriate interpretations and subjectivity, the lack of widespread access to robust solutions for digitizing the glass slides hindered the use of computerized analysis techniques.

1.2 Whole-slide imaging

Whole-slide imaging is the process of digitizing glass slides resulting in Whole-slide Images (WSI). WSIs are giga-pixel images that are generally stored in a multi-resolution pyramid structure. These files contain multiple downsampled versions of the original image (see Figure 1.3). Each image in the pyramid is stored as a series of tiles, to facilitate rapid retrieval of subregions of the image. Reading these images using standard image tools or libraries is a challenge because these tools are typically designed for images that can comfortably be uncompressed into RAM or a swap file. In addition, there is currently a lack of a universally accepted WSI format[4]. Different manufacturers have different proprietary WSI file formats such as SVS (by Aperio Technologies, USA) and NDPI (by Hamamatsu Photonics, Japan). OpenSlide[5] is a C library that provides a simple interface to read WSIs of different formats and paves the way for development of computerized algorithms that can operate at the WSI level.

Recent advances in slide scanning technology and cost reduction in digital storage capacity are enabling full digitalization of the microscopic evaluation of stained tissue sections ('digital pathology'). The advantages of these developments are many[7]: tele-education, remote diagnostics, immediate availability of archival cases, easier consultations with expert pathologists, and the possibility for computerized or computer aided diagnostics[8].

The accessibility provided by digital sharing of WSIs can improve the quality and efficiency of pathological services, primarily through Internet-based telepathology and teleconsultation. Telepathology is the practice of pathology over long distances by interpreting whole-slide images (WSI). Despite the managerial and legal barriers associated with the acquisition and use of telepathology in health care organizations and its high set up costs, the persistent shortage of pathologists and growing

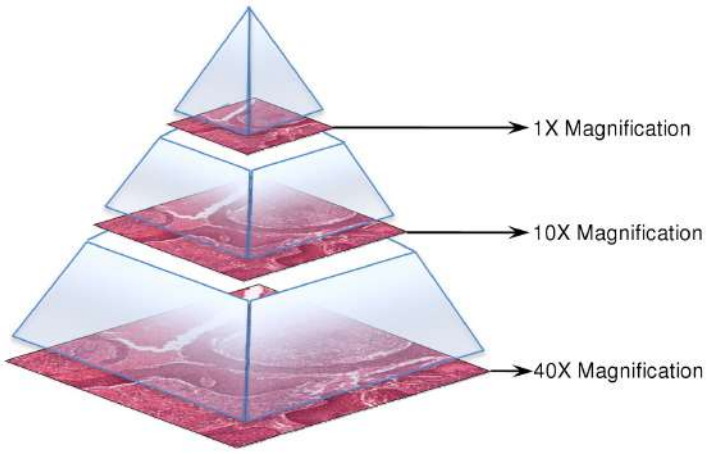

Figure 1.3: An illustration of how digital slides are stored in a pyramid structure. The image was produced by Yinhai Wang et al.[6]

demand for diagnostic consultation has fueled the adoption of telepathology and teleconsultation as a novel source of international trade[9].

But perhaps the most significant benefit of the digital revolution in pathology diagnostics is the possibility for computerized or computer-aided diagnostics. Once we are capable of having a computer assist a pathologist in the interpretation of WSIs, digital pathology will realize its full potential.

1.3 Computer-aided diagnosis for histopathology

Computer-aided detection (CADe) and computer-aided diagnosis (CADx) refer to computerized procedures which assist doctors in the interpretation of medical images. CADe and CADx are similarly represented as CAD. An example of such a system is one that automatically discriminates between malignant and benign WSIs of breast specimen. Application of CAD for histopathology can be broadly categorized into 3 major groups:

1. Detection, segmentation, and quantification of important tissue structures: The first use case for CAD in pathology addresses tedious routine diagnostic tasks that require great accuracy. Examples of these are 1) detection and segmentation of nuclei to analyze nuclear morphology as one of the hallmarks of cancerous conditions. 2) segmentation of glandular structures and analysis of the morphology of the gland as the key criterion for cancer grading (e.g. colon cancer). 3) detection of cancer metastases in lymph node sections.

2. Classification of histopathology imagery based on grade or lesion type: The classification of histopathology imagery is often the ultimate goal in many di-

agnostic tasks, particularly in cancer applications. Interpretation of pathology images for grading cancer or classifying lesions suffers from subjectivity. CAD may produce objective and accurate classification results to aid the decision of the pathologist. Examples applications are 1) Gleason scoring for prostate cancer, and 2) classification of breast proliferative lesions into benign, DCIS, and invasive cancer.

3. Disease diagnosis and prognosis: Computerized tools may yield relevant information for diagnosis and prognosis of diseases based on the assessment of subtle sub-visual changes in the patterns of important structures in histopathological images that are invisible to or hard to recognize for human vision. This can potentially lead to early diagnosis and prognosis of a disease and can, in turn, facilitate the subsequent clinical management of patients.

The requirement to design CAD systems for histopathology that can be used in clinical practice necessitates the development of robust and high throughput algorithms that can operate at the WSI level. Analyzing histopathological images, in particular, WSIs is, however, complicated by several factors. Firstly, histological stains show a large variability in their color and intensity. Such variations can potentially hamper the effectiveness of quantitative image analysis. Secondly, tissue preparation and digitization usually generates lots of artifacts which causes challenges for automated analysis. Finally, the large size of a WSI as well as the large number of heterogeneous structures that need to be analyzed inside a WSI make development of an accurate and reliable CAD for histopathology a challenge. Throughout this thesis, we focused on the development of computer-aided or computerized diagnosis tools that can operate at the WSI level by overcoming these challenges; hence enabling the systems to be utilized in a practical clinical setting.

The underlying fundamental method that sits at the core of most CAD methodologies is machine learning. The ability of machine learning algorithms to discover and identify patterns and relationships between them from complex datasets has made them very popular for the development of computerized and computer-aided diagnostic systems. In the next section, we will introduce machine learning as a subfield of artificial intelligence that is concerned with the design and development of algorithms that allow computers to learn.

1.4 Artificial intelligence and machine learning

The quest for artificial intelligence (AI) began over 70 years ago, with the idea that computers would one day be able to think like humans. AI emerged as a genuine

science following a month-long conference at Dartmouth College in 1956[10], where the most eminent experts gathered to brainstorm on creating an artificial brain. This event happened after Turing published a landmark paper[11] in which he devised his famous 'Turing test' to prove machine intelligence.

In the years following the Dartmouth conference, impressive advances were made in AI. Several focal areas of research for AI emerged between the 1950s and the 1970s. The first examples of AI systems were heuristic search programs pioneered by Newell and Simon[12] that used common sense rules drawn from experience to solve problems. Samuel[13] implemented a checkers-playing program, which improved itself through self-play. This program was one of the first working instances of a machine learning system. 'Shakey' the robot was the first general-purpose mobile robot to be able to reason about its own actions. The robot greatly influenced AI techniques and launched the field of mobile robotics. Rosenblatt's Perceptron[14], a computational model based on biological neurons, was a significant work that became the basis for the field of artificial neural networks. In 1980s, expert systems dominated AI. An expert system (also known as production system or rule-based system) is a computer system that emulates the decision-making ability of a human expert by coding into the computer a set of rules that determine what the computer should do on various inputs[15]. Expert systems were severely limited in their capacity to scale up, were expensive to maintain in industrial settings and had to be frequently updated (they did not 'learn').

Machine learning grew out of the quest for artificial intelligence and started to flourish in the 1990s. Machine learning is the science of giving "computers the ability to learn without being explicitly programmed" (Arthur Samuel, 1959)[13]. Nowadays, machine learning is applied in various fields such as self-driving cars, speech recognition, cyber security, healthcare, etc.

The machine learning paradigm can be viewed as "programming by example". Depending on the nature of data and the objective of the study, machine learning algorithms can be broadly divided into three categories− supervised learning, unsupervised learning, and reinforcement learning.

- Supervised learning: Given a database of training examples with a specific target label (property) in the form $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ where x_i denotes the feature vector of the ith example and y_i is its label, the goal is to construct a model $g : X \rightarrow Y$ that can accurately predict the label Y for future instances of data X. When this target property is a continuous real value, the task is referred to as regression. Otherwise, when the target property is a finite set of discrete values, the task is referred to as classification.

- Unsupervised learning: Aims to discover implicit relationships in a given unlabeled training dataset. The most common unsupervised learning is clustering which automatically partitions the data into groups of similar items called clusters.

- Reinforcement learning: In this type of learning, an agent interacts with its environment and learns its behavior based on the feedback it receives from its environment. Reward feedback known as reinforcement signal is required for the agent to learn its behavior. The agent must learn to act so as to maximize the total expected reward.

While reinforcement learning is frequently used for robotics, gaming and navigation, supervised and unsupervised learning algorithms are the more common forms of machine learning in medical imaging applications. Machine learning has been widely applied to medical imaging, perhaps most notably in the areas of computer-aided diagnosis (CAD). CADs using machine learning algorithms have been very successful in various fields, such as mass detection in breast cancer screening[16,17], nodule detection in lung cancer screening[18,19], detection of diabetic retinopathy[20,21], detection of white-matter hyper-intensities[22–24], and detection of various diseases in histopathology[8,25].

Conventional machine learning techniques for vision tasks were, however, limited in their ability to process images in their raw format. Traditionally, developing a machine learning system for processing image data required domain-specific expertise to carefully engineer quantitative features that could extract relevant visual characteristics from images to create a representation from which a learning algorithm (e.g. a classifier) could operate on. Deep learning, a specific subtype of machine learning, in contrast, can automatically discover the representations needed for detection or classification allowing the system to directly map an input image to an output vector.

The present thesis focuses on the development of various machine learning and image analysis techniques for assessment of histopathological WSIs of breast tissue. A large part of this thesis uses deep learning for analysis of histopathology images. We will provide more details on this increasingly important form of machine learning in the next section.

1.5 Deep neural networks

A standard artificial neural network (ANN) is a computing system made up of a number of simple, highly interconnected processing elements called neurons ar-

ranged in a series of layers, each producing a sequence of real-valued activations. Some of these neurons, known as input neurons, are designed to receive various forms of information from the environment, other neurons get activated through weighted connections from previously active neurons; these are known as hidden neurons. Output neurons are those that signal how the network responds to the information it has learned. A neuron is a unit that computes the dot product of the inputs x and their connection weights ω and applies an activation function $f()$ to generate the output in the form: $Y = f(\sum_i^n \omega_i x_i)$. Common choices for an activation function are the sigmoid function with the mathematical form of $\delta(x) = 1/(1 + e^{-x})$, tanh function, and rectified linear unit (ReLU) which computes the function $f(x) = max(0, x)$. Learning is about finding the weights for the network that makes it solve a specific task in some optimal sense and is achieved by general-purpose learning procedures (such as the backpropagation algorithm). ANNs are mostly fully connected, meaning that each hidden neuron is connected to every neuron in the previous and next layers.

These networks are called neural because they are loosely inspired by neuroscience. More broadly, the term neural networks evokes a particular paradigm for describing the networks of biological neurons that constitute the nervous system of mammalian brains. Shallow ANNs have been around for many decades[14,26] and consist of a small number of layers. Nonlinear activation functions in hidden units enable a shallow ANN (e.g. a single hidden layer) to approximate any continuous function (universal function approximators) provided that the network is given enough hidden units[27]. A deep neural network has many of these hidden layers. The advantage of a deeper network is in that it can represent many complex functions with fewer neurons and weights than shallow networks[28]. Although every problem is different in nature and the necessity for a deeper network architecture is an active area of research, empirical results for many real-world problems show that it is difficult to train shallow networks that are as accurate as the deeper counterparts[29].

Many neural network architectures have been developed for specic tasks. Convolutional neural networks are the specialized architectures for visual recognition tasks.

1.5.1 Convolutional neural networks

Convolutional neural networks or CNNs are a specialized kind of neural network that use convolution operators in place of matrix multiplication in at least one of their layers. CNNs are particularly designed for processing data that has a known, grid-like topology and have been very successful in image recognition tasks[30]. Unlike

fully connected networks, where the output of each neuron depends on the entire preceding layer, the output of neurons in convolutional layers only depend on a small sub-region of the input. Hidden neurons at different locations in the activation response share the same weights. This makes CNNs extremely parameter efficient and translation invariant, which are key in achieving better generalization in vision problems.

CNN layers typically consist of three stages. In the first stage, several convolutions are performed over the input image with a set of filters (local detectors), in parallel. This produces a set of linear activations. In the second stage, these activations are run through a non-linearity function (e.g. ReLU) to produce feature maps. In the third stage, a pooling layer may be inserted. Its function is to reduce the size of the feature maps and, therefore, to reduce the amount of parameters and computation in the network. Pooling operators on small windows of the input and computes a single summary statistic, e.g. maximum or average, for each window. The most common form of pooling layer is max pooling by a filter size of 2×2 and a stride of 2. This pooling layer reports the maximum output within a rectangular neighborhood of size 2×2 and downsamples the feature map by a factor of two along its width and height. Apart from computational benefits, pooling confers a local tolerance to spatial translations of the input, making the network robust to small translations of the input. Figure 1.4 shows an example of a simple CNN architecture.

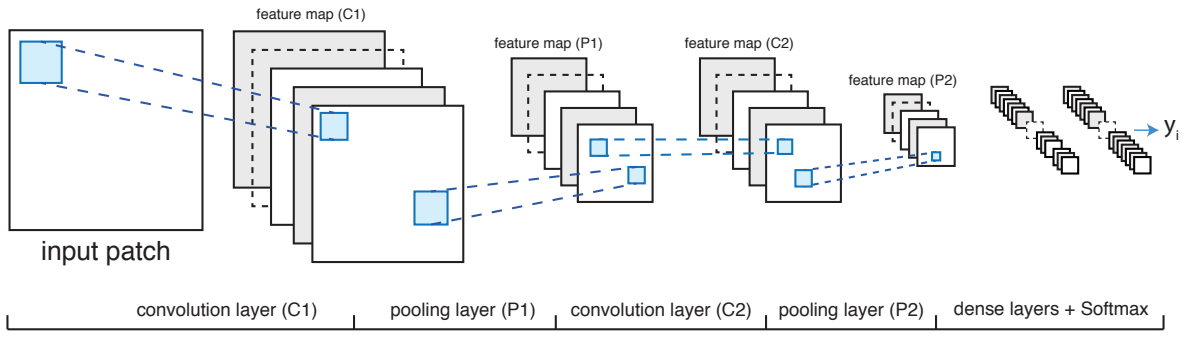

Figure 1.4: Graphical depiction of a simple CNN architecture with two convolutional and 2 fully connected layers.

CNNs have played an important role in the history of deep learning. The CNN built by Krizhevsky et al.[31] that won the ImageNet object recognition challenge, is among the most influential examples of computer vision. CNNs[32–34] have also been very successful in diagnostic imaging tasks and have won contests such as the ICPR 2012 and Miccai 2013 Contests on Mitosis Detection in Breast Cancer Histological Images, Segmentation of Neuronal Structures in EM Stacks Challenge, and IEEE ISBI 2016 challenge on the detection of cancer metastases in sentinel lymph nodes[35].

CNNs are well suited for analysis of medical images[36] and have a great potential to perform automatic lesion detection/segmentation, suggest differential diagnoses, and predict patient outcomes.

1.6 Aims and objectives

Application of machine learning to WSI is a promising yet largely unexplored fielld of research. The primary aim of the research described in this thesis was to develop automated systems for analysis of H&E stained breast histopathological images. This involved automatic detection of ductal carcinoma in-situ (DCIS), invasive, and metastatic breast cancer in whole-slide histopathological images. A secondary aim was to identify new diagnostic biomarkers for the detection of invasive breast cancer. To this end the research was undertaken with the following objectives:

1. Development of an algorithm for standardization of H&E stained WSIs (chapters 2 and 3);
2. Detection, classification and segmentation of primary breast cancer (chapters 4, 5 and 6);
3. Evaluation of the state of the art of machine learning algorithms for automatic detection of lymph nodes metastases (chapter 7);
4. Identifying and leveraging new stromal biomarkers to improve breast cancer diagnostics (chapter 8).

1.7 Thesis outline

This thesis is organized as follows:

Chapter 2 This chapter presents data on the sources of variation of the widely used H&E histological staining, as well as a new algorithm to reduce these variations in digitally scanned tissue sections.

Chapter 3 This chapter presents a fully automated algorithm for standardization of whole-slide histopathological images to reduce the effect of variations in the color and intensity of H&E stained histological slides that can potentially hamper the effectiveness of quantitative image analysis.

Chapter 4 This chapter presents a new algorithm for automatic detection of regions of interest in whole-slide histopathological images.

Chapter 5 This chapter presents and evaluates a fully automatic method for detection of ductal carcinoma in-situ (DCIS) in digitized H&E stained histopathological slides of breast tissue.

Chapter 6 This chapter presents context-aware stacked CNN for classification of breast WSIs into normal/benign, DCIS, and invasive ductal carcinoma (IDC).

Chapter 7 This chapter presents a comprehensive assessment of the performance of various machine learning systems in detecting metastases in H&E stained tissue sections of lymph nodes of breast cancer patients as part of the CAMELYON16 challenge. It also compares the performance of these algorithms to the pathologists in a diagnostic setting

Chapter 8 This chapter presents a new deep neural network framework for identifying tumor-associated stroma in WSIs of breast tissue biopsies as a highly discriminative diagnostic biomarker for detection of invasive cancer. It also presents results of an assessment for the potential of DCIS lesions to establish a microenvironment which supports invasive progression.

Chapter 9 This chapter summarizes the key contributions and findings of this thesis and discusses the implications of the work in details. In addition, the chapter draws final conclusions about the thesis and outlines the opportunities for future research in the area.

Stain variability assessment in histology slides

2

Babak Ehteshami Bejnordi, Nadya Timofeeva, Irene Otte-Höller, Nico Karssemeijer, and Jeroen AWM van der Laak

Original title: Quantitative analysis of stain variability in histology slides and an algorithm for standardization

Published in: Proceedings of SPIE Medical Imaging: Digital Pathology, vol. 9041, 2014

Abstract

This paper presents data on the sources of variation of the widely used hematoxylin and eosin (H&E) histological staining, as well as a new algorithm to reduce these variations in digitally scanned tissue sections. Experimental results demonstrate that staining protocols in different laboratories and staining on different days of the week are the major factors causing color variations in histopathological images. The proposed algorithm for standardizing histology slides is based on an initial clustering of the image into two tissue components having different absorption characteristics for different dyes. The color distribution for each tissue component is standardized by aligning the 2D histogram of color distribution in the hue-saturation-density (HSD) model. Qualitative evaluation of the proposed standardization algorithm shows that color constancy of the standardized images is improved. Quantitative evaluation demonstrates that the algorithm outperforms competing methods. In conclusion, this paper demonstrates that staining variations, which may potentially hamper usefulness of computer assisted analysis of histopathological images, can be reduced considerably by applying the proposed algorithm.

2.1 Introduction

Traditional diagnosis of cancer involves microscopic examination of histological slides acquired from tissue samples. Tissue sections are treated with multiple contrasting dyes to highlight different tissue structures and cellular features. Pathologists make diagnostic interpretations of the histology slides by assessing the cell structures and their spatial arrangement[37]. The task is laborious and time-consuming and prone to subjectivity[38,39]. Computer-aided diagnosis (CAD) can potentially address the issue of subjective interpretations by providing an objective quantitative assessment of digital pathology slides. CAD systems may facilitate cancer diagnosis by identifying abnormal areas on the slide and providing second opinions for patients[40]. However, variations in staining color and intensity complicate quantitative tissue analysis[40]. Such variations are due to inter-patient variation and inconsistencies in the preparation of histology slides (e.g. staining duration, stain concentration, tissue thickness). Although standardizing the staining protocols can minimize these effects, it is infeasible to remove all sources of variation[41] (e.g. tissue samples stained on different days of the week by following the same staining protocol in the same laboratory may result in different staining hue). Different approaches to overcome this problem have been proposed in the literature, based on the manipulation of the digital image after scanning of the slide. Methods based on normalization rely on extracting stain vectors and decomposing the image into individual stain components via color deconvolution[42]. The appearance of the images is normalized by adjusting the distribution of the color for each stain to a predefined range[43]. Methods based on color standardization match the color distribution of a histology image into a pre-defined template image by mapping its histogram-specific landmarks to the corresponding landmarks of the template image[44,45]. Previous work[45] has shown that separate standardization of the histogram for different tissue components yields an improved color constancy over global standardization approaches. However, the result of this standardization scheme relies on the accuracy of the tissue segmentation algorithm without considering the fact that pixels generally contain mixtures of stains. This study first investigates the contribution of several factors to the color variation of histology slides stained with the widely used hematoxylin and eosin (H&E) staining. Next, this study presents and evaluates a new algorithm for standardization of histology images based on an initial clustering of the image into two tissue components having different absorption characteristics for different dyes. Standardization of the color distribution is performed by aligning the 2D histogram of color distribution in the previously described hue-saturation-density (HSD) model[46].

2.2 Methods

2.2.1 Analysis of color variation in H&E stained histology slides

The H&E staining is the most commonly used staining technique in histopathology. Hematoxylin stains cell nuclei blue, and the counter-stain eosin stains cytoplasmic and non-nuclear components in different shades of pink[47]. An essential first step in the development of a CAD system is robust segmentation of the objects of interest (e.g. cell nuclei). The appearance of nuclei, however, can vary considerably among histological slides, which can cause significant adverse effects on the results of CAD. To investigate the extent and major reasons behind color variations in histological images, a technique was developed to quantify the distribution of color information. This technique applies the HSD model, which was specifically designed for absorption light microscopy[46]. The HSD model transforms RGB data into two chromatic components (c_x and c_y; which are independent of the amount of stain) and a density component (D; linearly related to the amount of stain). Expectation Maximization (EM) algorithm[48] was employed to estimate the parameters of a Gaussian mixture model in the $c_x c_y$ plane of the HSD transform. Each image is clustered by EM into three broad classes, representing different tissue components: nuclei, cytoplasm/stroma, and background. The probability density function for the ith class is represented by its mean μ_i and covariance matrix $\sum_i$. To understand the extent of the color variation of the hematoxylin stained tissue class, the variations in the mean value (i.e. perceived color) and the average eigenvalue (intra-specimen color variability) of the probability density function for the class representing the nuclei, is quantitatively described.

2.2.2 Standardization of histology slides

The proposed algorithm for standardization of histology images is based on an initial clustering of the image into three classes (nuclei, cytoplasm/stroma, background) in the $c_x c_y$ plane of the HSD transform. Background pixels (devoid of stain) show very low density values ($D < 0.2$). Given that the image background is white, all three RGB values have to exceed 180° to be classified as background. By application of singular value decomposition (SVD) on the remaining data points in the $c_x c_y$ plane, the two singular vectors with largest singular values are computed. The angle between each data point and the second singular vector (perpendicular to the first) was calculated. All the data points having an angle below 90° were classified as pixels absorbing the majority of hematoxylin stain and the remaining as pixels absorbing the majority of eosin stain. Figure 2.1 shows the scatter plot of the classified pixels in

the $c_x c_y$ space.

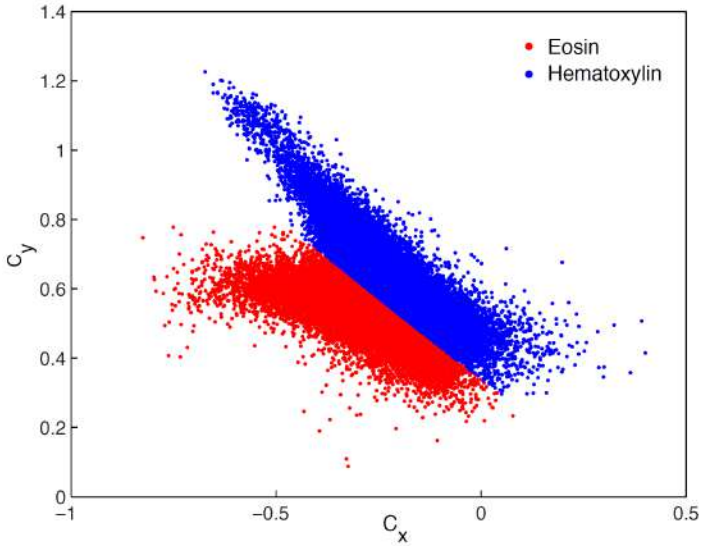

Figure 2.1: Scatter plot of all the pixels classified as hematoxylin and Eosin in $c_x c_y$ space by application of singular value decomposition.

The process for standardization of each tissue class i involves: (i) computing mean μ_i and covariance matrix $\sum_i$ of the data distribution for class i by modeling it as a Gaussian; (ii) centering the axis of the data points by translating the mean of the data distribution in class i to the point $(0, 0)$ and rotating the entire data distribution along the major eigenvector of $\sum_i$ to maximize the variance along the axes; (iii) computing landmarks $\{c_{min}, c_1, c_{median}, c_{99}, c_{max}\}$ for the rotated c_x and c_y values independently, where c_1, c_{median} and c_{99} denote the 1st, 50th and 99th percentiles of the data points belonging to class i, and $c_{\min}$ and $c_{\max}$ denote the minimum and maximum values of the entire distribution; (iv) interpolating the data distribution to match the predefined corresponding landmarks from the template image $\{t_{min}, t_1, t_{median}, t_{99}, t_{max}\}$ by performing piecewise linear mappings – from $[c_{min}, c_1]$ to $[t_{min}, t_1]$, $[c_1, c_{median}]$ to $[t_1, t_{median}]$, etc.; (v) rotating back the entire distribution along the major eigenvector of the corresponding tissue class in the template image; (vi) translating the entire distribution to the point (c_{xm}, c_{ym}) where c_{xm} and c_{ym} are the mean c_x and c_y values of the data distribution for the corresponding tissue class in the template image.

A weighted sum of the transformed $c_x c_y$ values for each class yields the final $c_x' c_y'$ for the entire distribution. The weight function is computed by first estimating the distribution of the two classes as two Gaussian components $k \in \{1, 2\}$ and then calculating their corresponding posterior probabilities for each observation X (excluding background observations) in the original $c_x c_y$ plane through $P(K|X) = \frac{\pi_k N(X|\mu_k, \sum_k)}{\sum_{j=1}^{k} \pi_j N(X|\mu_j, \sum_j)}$, where π_k is the prior probability and $N(X|\mu_k, \sum_k)$ is the probability

density function of component K for the observation X. This yields the following transformation:

$$c_x^{'}c_y^{'} = c_{x_1}c_{y_1} \times P(k=1|c_{x_1}c_{y_1}) + c_{x_2}c_{y_2} \times P(k=2|c_{x_2}c_{y_2}) \tag{2.1}$$

where $c_{x_1}c_{y_1}$ and $c_{x_2}c_{y_2}$ are the transformed c_xc_y values belonging to each class, and $c_x^{'}c_y^{'}$ is the yielded final transformation. After scaling the density histogram of the image to the density histogram of the template image using 10 evenly spaced percentiles as landmarks, the RGB channels are reconstructed by the reverse HSD transform[46].

Histology images

The image data used in this study originate from a set of 45 digitized H&E stained histopathology slides of lymph nodes from three different patients. The slides were stained in three different laboratories on different days of the week (15 slides for each lab). All the slides were digitized using a CCD RGB camera mounted on a light microscope with a $40\times$ objective lens (Olympus dotSlide system, Olympus, Japan). Three representative regions of interest (ROI) images were acquired from each slide yielding a total of 135 images. Each image is of size 2300×3300 pixels with square pixels of size $0.16\mu m$ in the microscope image plane. Figure 2.2 shows three sample images stained in different laboratories.

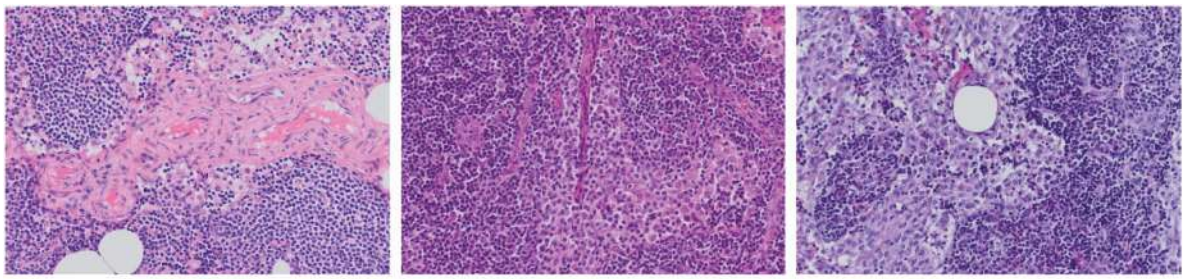

Figure 2.2: Sample images from the three laboratories. (a), (b), and (c) are three sample ROI images stained by laboratory 1, 2 and 3, respectively.

2.2.3 Experiments

To evaluate the major factors causing stain variations in histology images and to understand the extent of these changes and also to evaluate the performance of the proposed color standardization algorithm two experiments were performed.

The aim of the first experiment was to statistically measure the variations in the mean (c_x, c_y) values (describing absolute hematoxylin color) and the average eigenvalues (describing intra-slide hematoxylin variation) of the probability density function for the class representing the nuclei in the $c_x c_y$ plane of the HSD model. The importance of three potential factors that contribute to staining variations including patient, staining laboratory and staining day of the week was studied. Provided that the staining solutions used by the automated staining machines is changed once or twice per week, staining on different days of the week may result in a stain having different spectral characteristics. Multivariate analysis of variance (MANOVA)[49] with Pillai's statistics was used to determine the statistical significance of differences between these variables.

In the second experiment, the performance of the proposed standardization algorithm is both qualitatively and quantitatively evaluated. Qualitative evaluation is performed by visualizing the original image and assessing the visual color constancy of the standardized image with respect to the template image. For quantitative evaluation of the results, the color constancy of the nuclei segmented regions were evaluated and compared to the global standardization (GS) algorithm[44] and the appearance normalization method[43] by calculating the normalized median intensity (NMI)[45,50]. The standard deviation of the NMI values and the coefficients of the variations (i.e. standard deviation divided by mean) were computed for all the images in the dataset before and after standardization using the proposed algorithm. The nuclei were segmented by imposing a threshold of $T = 120$ (determined empirically) on the average of the RGB values for each image.

2.3 Results

Experiment 1: MANOVA results show that differences between labs may be considerable, and that for certain labs staining results differ significantly between days of the week (Table 2.1). Comparable results were found for the mean (c_x, c_y) values and the average eigenvalues of the nuclei. No significant color variations were observed between tissues from different patients.

Figure 2.3 shows boxplots of the average eigenvalue distributions of the nuclei segmented regions of the slides stained in different laboratories on different days of the week. The mean and standard deviation of the average eigenvalues for different days of the week are considerably higher for the third laboratory.

Experiment 2: Figure 2.4 shows the result of the standardization by the proposed algorithm. The original images in Figure 2.2b and Figure 2.2c are standardized by using the image in Figure 2.4a as the template image. The standard deviation (SD)

Table 2.1: Multivariate analysis of variance with Pillai's statistics. L1, L2, and L3 denote Laboratories 1 to 3.

Variables	c_x, c_y	Average eigenvalues
Patients	generally no significant effect	no effect
Laboratory	significant differences between L3 and L1/L2 ($p < 0.005$) weak difference between L1 and L2	significant differences between L3 and L1/L2 ($p < 0.005$) weak difference between L1 and L2
Days of the week	L1 ($p < 0.005$) L2 not significant L3 ($p < 0.005$)	L1 ($p < 0.02$) L2 not significant L3 ($p < 0.005$)

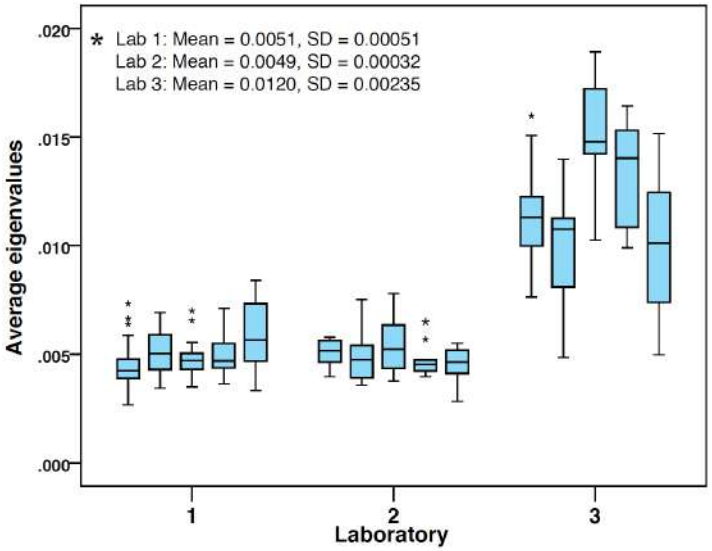

Figure 2.3: Boxplots of average eigenvalue distributions for the three laboratories for different days of the week. The mean and standard deviation have been calculated for the five weekday distributions for each laboratory.

of the normalized median intensity and its coefficient of variation (CV) for the segmented nuclei are presented in Table 2.2 for the three laboratories. The proposed method outperforms GS algorithm and the appearance normalization method by yielding the lowest SD and CV of normalized median intensity.

2.4 Discussion and conclusion

This paper investigated the effect of several factors on the color variation of H&E stained histology slides. The experimental results demonstrate that staining protocols in different laboratories and staining on different days of the week are the major factors causing color variations in histopathology images. This paper also presented a new algorithm for reducing stain variations in histology slides. Qualitative assess-

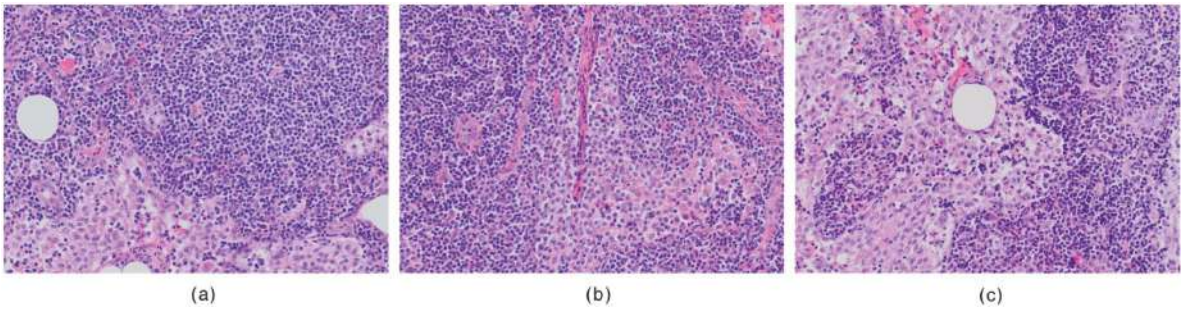

Figure 2.4: Standardization of H&E stained histopathology images. (a) The template image used for extracting the parameters of the standardization algorithm. (b,c) The standardization results for images shown in Figure 2.2(b,c).

Table 2.2: Standard deviation and coefficient of variation of NMI for all the images in the three laboratories.

Algorithm	*Laboratory 1*		*Laboratory2*		*Laboratory 3*	
	NMI SD	NMI CV	NMI SD	NMI CV	NMI SD	NMI CV
Original	0.0363	0.0450	0.0260	0.0317	0.0367	0.0484
Macenko[43]	0.0251	0.0275	0.0207	0.0228	0.341	0.4520
Global standardization[44]	0.0240	0.0277	0.0203	0.0238	0.0219	0.0256
Proposed algorithm	**0.0068**	**0.0081**	**0.0077**	**0.0093**	**0.0094**	**0.0123**

ment of the results show the efficacy of the algorithm in maintaining color constancy of the histology images. The empirical results show that the proposed algorithm outperforms the global standardization algorithm and the appearance normalization method (based on automatic derivation of stain vectors) by yielding a lower standard deviation and coefficient of variation of the normalized median intensity. The algorithm presented in this paper can be applied to other histological stains and tissues. This will be the subject of future work.

Acknowledgments

The authors wish to acknowledge the financial support by the European Union FP7 funded VPH-PRISM project under grant agreement n°601040. We also gratefully acknowledge financial support from the Stichting IT Projecten Nijmegen (NT) and the Maurits en Anna de Kock foundation for image analysis equipment.

Stain specific standardization of whole-slide histopathological images

3

Babak Ehteshami Bejnordi, Geert Litjens, Nadya Timofeeva, Irene Otte-Höller, André Homeyer, Nico Karssemeijer, and Jeroen AWM van der Laak

Original title: Stain specific standardization of whole-slide histopathological images

Published in: IEEE Transactions on Medical Imaging, 35(2), 404-415, 2016.

Abstract

Variations in the color and intensity of hematoxylin and eosin (H&E) stained histological slides can potentially hamper the effectiveness of quantitative image analysis. This paper presents a fully automated algorithm for standardization of whole-slide histopathological images to reduce the effect of these variations. The proposed algorithm, called whole-slide image color standardizer (WSICS), utilizes color and spatial information to classify the image pixels into different stain components. The chromatic and density distributions for each of the stain components in the hue-saturation-density color model are aligned to match the corresponding distributions from a template whole-slide image (WSI). The performance of the WSICS algorithm was evaluated on two datasets. The first originated from 125 H&E stained WSIs of lymph nodes, sampled from 3 patients, and stained in 5 different laboratories on different days of the week. The second comprised 30 H&E stained WSIs of rat liver sections. The result of qualitative and quantitative evaluations using the first dataset demonstrate that the WSICS algorithm outperforms competing methods in terms of achieving color constancy. The WSICS algorithm consistently yields the smallest standard deviation and coefficient of variation of the normalized median intensity measure. Using the second dataset, we evaluated the impact of our algorithm on the performance of an already published necrosis quantification system. The performance of this system was significantly improved by utilizing the WSICS algorithm. The results of the empirical evaluations collectively demonstrate the potential contribution of the proposed standardization algorithm to improved diagnostic accuracy and consistency in computer-aided diagnosis for histopathology data.

3.1 Introduction

Histopathology involves microscopic examination of stained histological slides to study presence and characteristics of disease. Tissue sections are stained with multiple contrasting dyes to highlight different tissue structures and cellular features[37]. This staining provides invaluable information to the pathologists for diagnosing and characterizing various pathological conditions. Pathologists make diagnosis of the disease based on features such as morphology and spatial arrangement of cells[40]. This task is, however, laborious and prone to subjectivity[38,39]. Several previous studies have shown poor concordance among pathologists in histopathological grading of prostate, cervical, and breast cancer[51–53]. Computer-aided diagnosis (CAD) can potentially alleviate shortcomings of human interpretation. CAD can facilitate diagnosis by sieving out obviously benign slides and providing quantitative characterization of suspicious areas.

The appearance of the histological stains (e.g. the widely used hematoxylin and eosin (H&E) staining) often suffer from large variability[54]. While pathologists can effectively cope with staining variability, the performance of CAD systems can be hampered by color and intensity variations. Such variations in digital pathology images may be attributed to a number of factors, including specimen preparation and staining protocol inconsistencies (e.g. temperature of solutions), variations in fixation characteristics, inter-patient variation, and the scanner used to digitize the slides[54]. The use of standardized staining protocols and automated staining machines may improve staining quality by yielding a more accurate and consistent staining. However, eliminating all the underlying sources of variation is infeasible[41]. The problem is more acute in studies in which different laboratories share digital images. As an alternative, algorithms for automated standardization of digitized whole-slide images (WSI) have been published[41,43,45,55–59] (described in detail below). Ideally, such an algorithm takes as input an arbitrary WSI and yields a normalized version of the image, with standardized appearance of the different dyes. The algorithm should be able to eliminate all sources of variation.

Many color standardization algorithms, also referred to as color normalization algorithms, are based on stain-specific color deconvolution[42]. Stain deconvolution requires prior knowledge of reference stain vectors for every dye present in the WSI. Ruifrok *et al.*[42] suggested a manual approach to estimate color deconvolution vectors by selecting representative sample pixels for each stain class. A similar approach was used in[55] for extracting the stain vectors. Manual interaction for estimation of stain vectors, however, strongly limits its applicability in large studies. The method devised by Macenko et al.[43] enables automatic extraction of reference stain vectors

by finding the fringe of pixel distributions in the optical density space. The method, however, yields poor estimation of the stain vectors in the presence of strong staining variations. Several other approaches have been suggested for automatic extraction of stain vectors[41,56]. The major drawback of these methods is that the estimation of the stain vectors relies solely on the color information present in the image. The outstanding ability of a pathologist to identify stain components is not only because of using the color information but also because of incorporating the spatial dependency of tissue structures (e.g. cell nuclei have a near elliptic shape and in H&E staining are mostly stained with hematoxylin while other tissue structures are mostly eosin stained). Neglecting to take into account the spatial dependency of tissue structures for determination of the stain vectors significantly limits the robustness of these methods in presence of severe staining variations.

Reinhard *et al.*[57] proposed one of the first techniques to standardize image colors with respect to a template image through the use of a color model. Their proposed technique aligns each of the color channels of the Lab color model[60] to the corresponding channels of a predefined template image. However, the use of a single transformation function for each channel will, in practice, rarely suffice. In general, dyes have independent contribution to the final color in the image, as each dye has its own specific reaction pattern. Consequently, using a single transformation function may lead to improper color mapping of the stain components in the standardization process. This problem can be addressed by applying separate transformations to different stain classes[55,58] or different tissue classes[45]. The applicability of the approach described in[45] is limited to image patches that necessarily contain all the considered tissue classes. Moreover, the use of Gaussian mixture model in[45] and[55] for segmentation of tissue or stain classes lacks robustness to strong staining variations. The performance of the algorithm described in[58] also decreases when there is a considerable imbalance in the amount of different stain classes in the image. This is mainly because the accuracy of the stain classification approach utilized (based on singular value decomposition) decreases as the imbalance between the amount of different stain components increases.

In addition to the limitations discussed above, the majority of these studies have 3 major shortcomings which limit their applicability to studies using large cohorts. First of all, the efficacy of most of the algorithms was not evaluated on data from multiple laboratories (preferably both academic and non-academic) with different staining protocols, which potentially causes the most severe staining variations[58]. Secondly, most of the proposed algorithms to date have been solely focusing on standardization of patch images containing a small region within the WSI. Development of a fully automated CAD system for large-scale digital pathology requires

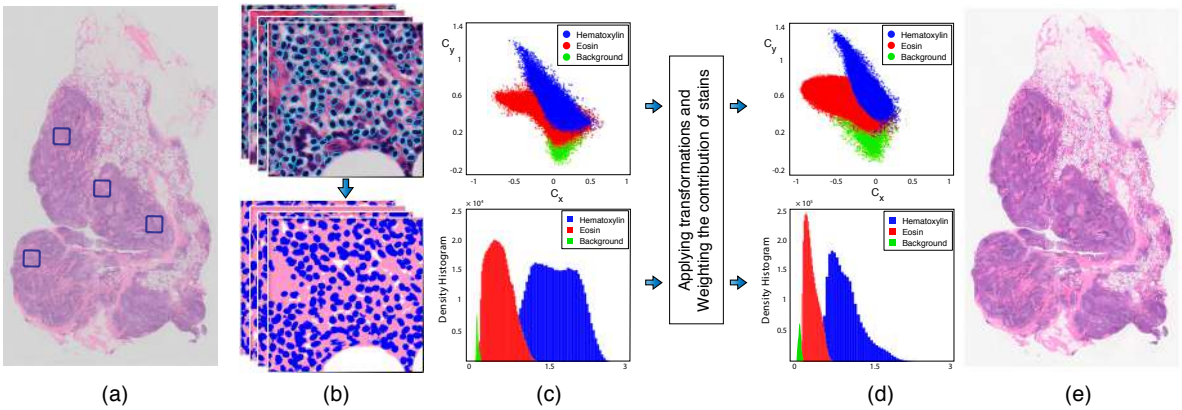

Figure 3.1: Illustration of selected steps of the WSICS algorithm. (a) Original WSI of a lymph node tissue section. Several regions containing tissue are randomly selected for automatic extraction of representative samples for the hematoxylin, eosin and background class. (b) The selected regions of interest are classified into: pixels absorbing mostly hematoxylin, pixels absorbing mostly eosin, and background pixels. (c) The chromatic distribution and density histogram of pixels are defined in the HSD model to be transformed to match a template WSI. (d) The result of transforming the chromatic and Density distributions after application of weights. (e) Reconstructing the RGB image by applying reverse HSD transform using the standardized chromatic and density components.

algorithms that can deal with WSI data. Whole-slide extension of the existing patch-based standardization algorithms may not be straightforward and requires automatic extraction of training patches from the entire slide to define the required color transformations. In[59] an automated algorithm has been proposed for WSI standardization which can handle variations caused by using different WSI scanners. However, variations caused by other sources such as staining protocols are more problematic[58] and can not be dealt with by this approach. Finally, although the major aim of most published algorithms is to enhance the performance of CAD systems, the efficacy of these algorithms was typically not evaluated on an existing CAD system. To the authors' knowledge no algorithm has been proposed to date that addresses WSI standardization in a fully automated manner in presence of all sources of variation.

This paper presents and evaluates a novel fully automatic algorithm for standardization of whole-slide H&E stained histopathological images. The algorithm, called whole-slide image color standardizer (WSICS), is based on transformation of the chromatic and density distributions for each individual stain class in the hue-saturation-density (HSD) color model[46]. While standardization strongly facilitates accurate recognition of stain components, a good technique for determination of stain components can help in developing a robust standardization algorithm. Such a

technique should recognize different dye components in presence of various sources of stain variation. Unlike the available supervised and unsupervised pattern recognition techniques for dye recognition in the literature, which rely solely on color features, the proposed algorithm makes use of spatial information, making it robust against severe color and intensity variations. Standardization of the stain components in the WSI is achieved by aligning the chromatic and density distributions of the stain components to predefined corresponding histograms from a template WSI. The WSICS algorithm is compared to three state-of-the-art standardization algorithms both qualitatively and quantitatively. Empirical evaluation of the impact of our standardization algorithm on the performance of an already published CAD system for necrosis quantification[61] is also presented.

3.2 Methods

3.2.1 Overview of the Proposed Whole-slide standardization Method

The WSICS algorithm takes as input a WSI and outputs a standardized image with staining characteristics similar to a predefined template WSI. We interface our standardization algorithm with a 2D WSI, using our own developed open source library[62] which is built on top of the OpenSlide library[63]. This library allows us to read a large set of WSI file formats (e.g. .tif, .vsi, .mrxs, .svs, etc.).

Our proposed algorithm initially classifies the pixels in the WSI into different dye classes and then applies standardization on the chromatic and density components of the HSD model[46]. Figure 3.1 presents an overview of the WSICS algorithm. The proposed algorithm has 6 basic steps:

1. Applying HSD transform.

2. Automatic extraction of samples for the hematoxylin, eosin and background classes from the WSI and deriving the chromatic and density distributions of these classes.

3. Transforming the $2D$ chromatic distribution for each dye class to match the chromatic distribution of the corresponding class from a template slide.

4. Transforming the density distribution for each dye class to match the density distribution of the corresponding class from a template slide.

5. Weighting the contribution of stains for each pixel and obtaining final chromatic and density transformations.

6. Applying inverse HSD transform.

Detailed description of the WSICS algorithm's steps are discussed below.

3.2.2 HSD transform

The algorithm first applies the hue-saturation-density (HSD) color transformation[46]. In[46] we showed that the HSD model is better-suited for analysis of transmitted light microscopy compared to the RGB and HSI models. The HSD model transforms RGB data into two chromatic components (c_x and c_y; which are independent of the amount of stain) and a density component (D; linearly related to the amount of stain). Theoretical RGB intensities obtained from varying stain densities should result in a single point in the $c_x c_y$ chromaticity plane of the HSD transform[46]. However, in practice, the use of broad-band camera filters and existence of pixel inhomogeneity (stain variability over the specimen area occupied by a pixel) lead to dispersion of chromatic data in this plane for each dye component. As a result, the chromatic data of the pixels stained with a particular dye component will form a distribution, which is represented by $F(c_x, c_y)$. In H&E staining the chromatic distribution of the hematoxylin stain $F_H(c_x, c_y)$, and the chromatic distribution of the eosin stain $F_E(c_x, c_y)$ have a significant overlap.

3.2.3 Deriving the chromatic and density distributions of hematoxylin, eosin, and background

WSIs are generally stored in a multi-resolution pyramid structure. Image files contain multiple downsampled versions of the original image. Each image in the pyramid is stored as a series of tiles, to facilitate rapid retrieval of subregions of the image which enable us to quickly identify regions that are rich in tissue.

For a WSI, extraction of representative samples for hematoxylin, eosin and background classes starts with identifying the tiles containing more than 75% of tissue (non-background pixels) on the lowest magnification. The pixel size of the image at this magnification is $3.88\mu m \times 3.88\mu m$. Each tile is a 64×64 pixel image. A pixel is classified as background if its overall density is lower than 0.2.

Next, we randomly select one of the tiles identified in the lowest magnification and apply Restricted Randomized Hough Transform[64] to detect candidate nuclei in the corresponding area in the highest magnification. If the number of detected nuclei surpasses a predefined threshold (200 nuclei) the image is classified into different dye components, and the labeled pixel samples are stored. This process is repeated for

a large number of randomly selected tiles until a predefined number of samples (in this study 3 million pixels) are acquired for each class.

To make the sampling process robust against the inhomogeneities present in some slides, it is possible to sample less pixels in each tile image but instead sample from more random patches from the entire WSI. This, however, comes with an additional computation cost.

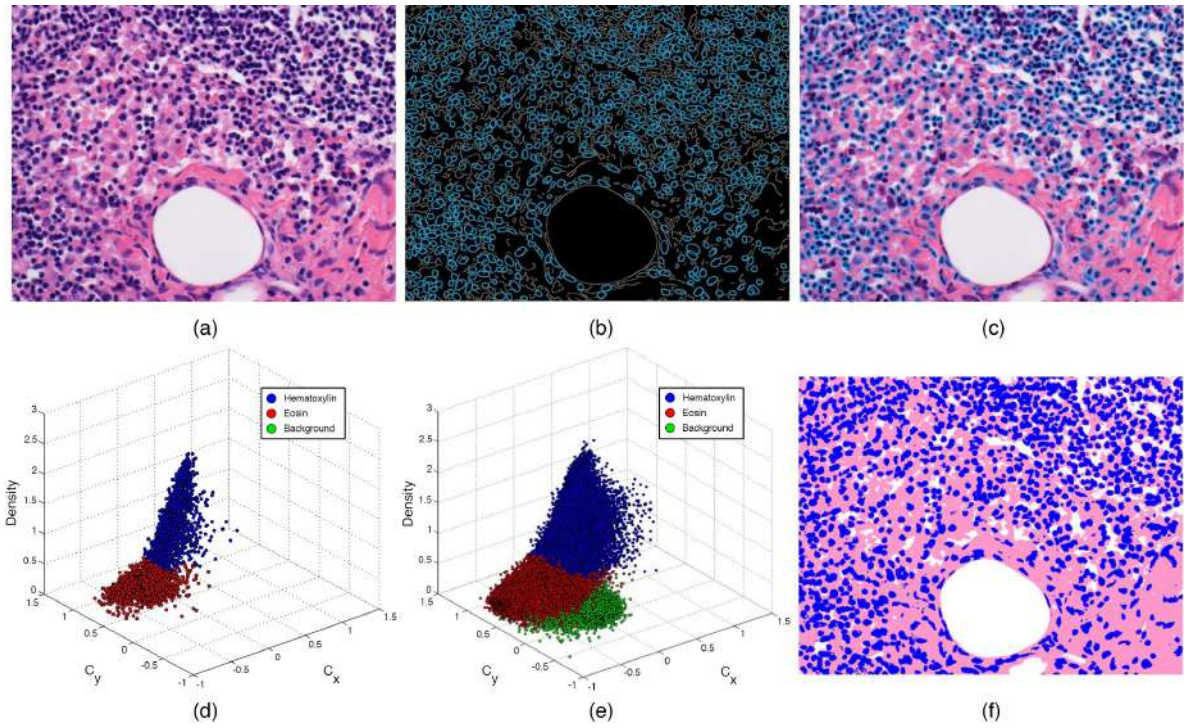

Figure 3.2: Illustration of the classification steps. (a) Sample image of a lymph node tissue section. (b) Utilizing Canny edge detector on the density image in conjugation with the Restricted Randomized Hough Transform to detect and estimate the boundaries of the nuclei. (c) The result of detected nuclei overlayed on the original image after artifact rejection. (d) Scatter plot showing c_x, c_y, and density features of samples extracted from hematoxylin and eosin class. (e) Scatter plot after classifying the entire image. Note that the data points associated with the background class have been obtained by thresholding. (f) The classification result produced by our method.

Classification of pixels into different dye components requires prior knowledge of $c_x c_y$ color vectors for every dye in the WSI. To enable fully automated classification, the algorithm automatically extracts training samples for each stain class (the class absorbing mostly hematoxylin and the class absorbing mostly eosin) from the image, thus obviating the need for manually labeled training data. Our technique for generating training samples makes use of prior shape knowledge (cell nuclei are usually ellipse shaped). Incorporating spatial information such as prior shape knowledge into the classification problem provides robustness to color and intensity

variations.

The process for classifying different tissue components is illustrated in Figure 3.2. The first step involves locating the nuclei and estimating their boundaries with ellipses by utilizing the Restricted Randomized Hough Transform[64] on the Canny edge detected image over the density component of the HSD model (see Figure 3.2b). To remove artifacts among the detected candidate nuclei, percentile thresholding is applied on the average optical density of the red camera channel (D_r) and the average overall density (D). The optical density of a channel is defined as:

$$D_{ch} = -\ln\left(\frac{I_{ch}}{I_{o,ch}}\right)$$

where I_{ch} is the intensity of channel ch (which can be R, G, or B in RGB color model), and $I_{o,ch}$ is the intensity of channel ch when no stain is present. The overall density (D) refers to the density component of the HSD model. Among the detected candidates, objects with very low average D (e.g. background) and objects presenting very low average D_r (e.g. red blood cells) are removed by applying these thresholds. We found the 8th percentile to be a suitable threshold for both average D_r and average D measures. A random selection of samples from the pixels belonging to the detected nuclei constructs conservative representative training samples for the tissue class mostly absorbing hematoxylin. As shown in Figure 3.2c, our algorithm does not require exhaustive detection of all the nuclei, but rather a proportion of the nuclei within the image.

In the next step, all the pixels with very low overall density ($D < 0.2$) are classified as background (devoid of stain). Provided that background pixels are white, the optical density of the red, green, and blue camera channels of the pixel should each be lower than 0.25 to be classified as background.

Training samples for the eosin class are obtained by first removing all the background pixels and the candidate pixels for the hematoxylin class (before artifact rejection) and then applying 5th percentile thresholding on the D_r of the remaining pixels. Note that the training samples for the eosin class are randomly selected from these pixels and are equal in number to the training samples from the hematoxylin class. The pixels selected during the sampling process will serve as ground truth data for a classifier. By applying a threshold on the optical density of the red channel, we deliberately avoid sampling weekly stained pixels in the eosin class and try to retain samples that are as clean as possible.

Finally, a binary k-NN classifier ($k = 7$) is trained using the extracted samples from the hematoxylin and eosin class. The chromatic information (c_x and c_y) and the density of each pixel in the HSD color model are used as classification features. The class labels for all the remaining pixels in the randomly selected training patches are

predicted using the trained classifier, yielding the final pixel classification result.

The derived chromatic and density distributions for each of the stain classes (as shown in Figure 3.2e) are used for subsequent chromatic and density transformations.

3.2.4 Non-Linear Transformation of chromatic information

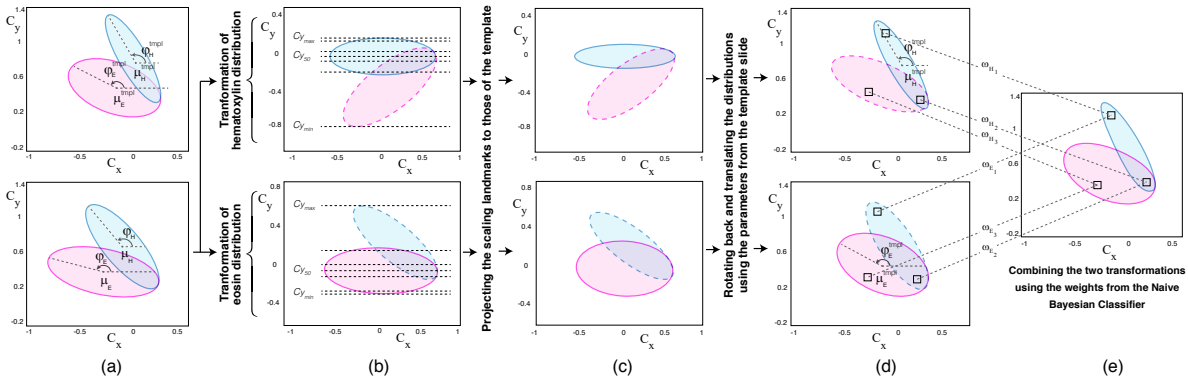

Figure 3.3: Illustration of the non-linear transformation of hematoxylin and eosin chromatic distributions. Note that the transformation applied to the background class is not shown in this figure. (a) The chromatic distribution of hematoxylin (blue) and eosin (pink) staining for the template slide (top left) and the slide to be standardized (bottom left). The angle (ϕ_i) and the mean (μ_i) of the H&E distributions are extracted, where $i \in \{Hematoxylin, Eosin\}$. (b) Extraction of the scaling parameters after translation of each distribution to the origin and rotation along ϕ_i. The extracted parameters comprise of the minimum, 1st, 25th, 50th, 75th, and 99th percentiles, and maximum of the projected values along each axis (for simplicity this is only shown for parameters along the c_y axis). (c) The result of scaling the hematoxylin and eosin distributions by using the scaling parameters derived from the template slide. (d) Rotation of the distributions along ϕ_i^{tmpl} and translating them to μ_i^{tmpl}. (e) The final transformation obtained using the weighted combination of the two transformations in *(d)*.

To align the chromatic distribution of each of the classified dye components to match the corresponding class distribution in a template slide, we perform $2D$ registration of the color information in the $c_x c_y$ chromatic plane.

Let $F_i(c_x, c_y)$ denote the chromatic distribution for class i in the image, where $i \in \{Hematoxylin, Eosin, Background\}$. We define the registration problem as finding the transformation function T_i such that:

$$F_i(T_i(c_x, c_y)) \sim F_i^{template}(c_x, c_y) \tag{3.1}$$

The procedure for registration of the chromatic distribution has two steps: (1) extraction of statistical parameters from the template slide and (2) aligning the $2D$ chromatic distribution. Each of the steps are described in the following subsections.

Template Parameter Extraction

Training samples for the two stain classes in the template image are obtained automatically as described in section A above. From these, 3 sets of parameters are calculated (two for the stain classes and one for the background class). These parameters are the mean μ_i^{tmpl}, angle ϕ_i^{tmpl}, and scaling landmarks of the $F_i^{template}$ distribution. Let $\Sigma_i^{tmpl} = cov(c_x, c_y)$ denote the covariance matrix of $F_i^{template}$. The angle ϕ_i^{tmpl} of this distribution with respect to the c_x axis is derived by calculating the major eigenvector of Σ_i^{tmpl}. To compute the scaling parameters, the entire $F_i^{template}$ distribution is translated to the origin. This is followed by a rotation step along ϕ_i^{tmpl} to maximize the variance along the c_x axis. Finally, the scaling landmarks are defined after projection of the rotated distribution onto each of the c_x and c_y axes, comprising the minimum, 1st, 25th, 50th, 75th, and 99th percentiles, and maximum of the projected values along each axis.

The only parameters for the background class are the mean values of the c_x and c_y channels. This is mainly because the white background class does not require major color standardization but rather density standardization which is achieved by adjusting its density component in the HSD model.

Alignment of $2D$ Color Histogram

At this step, the $2D$ chromatic distribution for each of the three classes in the WSI to be standardized are aligned with the chromatic distribution of the corresponding class in the template WSI. Figure 3.3 illustrates different steps for the non-linear transformation of chromatic information. Let $F(c_x, c_y)$ denote the chromatic distribution of the WSI to be standardized. We apply three separate transformations $T_i(c_x, c_y)$, each time focusing at standardizing the chromatic distribution of a particular class i.

Let $F_i(c_x, c_y)$ denote the chromatic distribution of the pixels belonging to class i sampled from the slide to be standardized. For different pixel classes in this slide, statistical parameters are extracted identical to those previously extracted from the template slide. The process for standardization of each stain class i starts with translating the entire $F(c_x, c_y)$ distribution by subtraction of the mean of $F_i(c_x, c_y)$ distribution and rotation along the major eigenvector of Σ_i, where Σ_i denotes the covariance matrix of $F_i(c_x, c_y)$ (see Figure 3.3a and Figure 3.3b). Then we apply piece-wise lin-

ear scaling to match landmarks from the current distribution to those of the template slide. The result is shown in Figure 3.3c. In the next step, the scaled distribution is rotated back along the ϕ_i^{tmpl} angle to be aligned with the major eigenvector of the the corresponding distribution in the template WSI. The final step in the transformation of the $F_i(c_x, c_y)$ involves translation of the distribution to μ_i^{tmpl}. The standardized chromatic distributions with focus on transforming hematoxylin and eosin classes are shown in the first and second row of Figure 3.3d, respectively.

The color transformation of the background class is yielded by subtracting μ_b from and adding μ_b^{tmpl} to the $F(c_x, c_y)$ distribution, where μ_b, and μ_b^{tmpl} denote the mean of the background class in $F(c_x, c_y)$, and $F^{template}(c_x, c_y)$ distributions, respectively. At the end of this step we have 3 separate transformation functions, one for each class.

3.2.5 Weighing the Contribution of Classes

Provided that in H&E staining, pixels may contain a mixture of stains, it is important to take into account the contribution of each stain for each pixel in our final transformation. Therefore, we define the final transformation as the weighted combination of the transformation functions associated for each class. To generate per pixel weights accounting for relative absorption of each stain, a naive Bayesian classifier is trained. The ground truth data for training the classifier originates from the automatically classified samples extracted in step 1 of the WSICS algorithm. By using the chromatic and density components of the HSD transform as features, we calculate the posterior probability of each pixel in the WSI belonging to each class. Finally, the weighted combination of the three transformations gives:

$$(c_x^{'}, c_y^{'}) = \sum_{i=1}^{3} \omega_i * T_i(c_x, c_y) \tag{3.2}$$

where $i \in \{Hematoxylin, Eosin, Background\}$ and ω_i denotes the weight for class i, and $(c_x^{'}, c_y^{'})$ denotes the final transformed chromatic vector in the HSD model. The result is shown in Figure 3.3e.

The naive Bayesian classifier was chosen to generate the membership degree of each pixel to each of the stain classes because of its straightforward probabilistic interpretation and its relative simplicity, requiring no parameter tuning in contrast to more complex classifiers. Given the nature of the classification task (3 features and millions of samples) a more complex classifier is not needed.

3.2.6 Density Standardization

The density component of the HSD color model was also transformed to match the density profile of the template image using a weighted combination of linear transformations with respect to each class. Applying linear transformation of the densities for separate classes without using the weights may lead to severe artifacts. The reason for this is that pixels lying in the vicinity of classification boundary but belonging to different classes will be standardized with separate transformation functions, hence transforming into new density values which might differ significantly. To address this problem, the weights generated by the naive Bayesian classifier in the previous section were used to create a smooth density standardization. In the first step a weighted mean and a weighted standard deviation were computed for each of the density distributions corresponding to the hematoxylin, eosin and background classes separately. For each of these classes, the distributions were standardized by matching the mean and standard deviation of the distribution to the corresponding class statistics in the template image. The transformed density distribution for class i is therefore determined by:

$$D_i^{'} = \frac{D_i - \mu_i}{\sigma_i} \times \sigma_i^{tmpl} + \mu_i^{tmpl} \tag{3.3}$$

where $i \in \{Hematoxylin, Eosin, Background\}$, μ_i and σ_i are the weighted mean and the weighted standard deviation of the density distribution for class i in the WSI to be standardized and μ_i^{tmpl} and σ_i^{tmpl} are the corresponding values in the template WSI. D_i and $D_i^{'}$ denote the densities before and after transformation.

By using the weights obtained from the naive Bayesian classifier, a weighted combination of the linear transformations for the three classes were computed to yield the final density transformation:

$$D^{'} = \sum_{i=1}^{3} \omega_i \times D_i^{'} \tag{3.4}$$

where ω_i denotes the posterior probability of class i, $D_i^{'}$ denotes the standardized density associated with the distribution of class i, and $D^{'}$ is the final standardized density component.

3.2.7 Inverse HSD transform

In the final step, the standardized chromatic components $c_x^{'}$ and $c_y^{'}$, and the standardized density component $D^{'}$ were used to get back to RGB model. This was achieved by following the reverse HSD transformation illustrated in[46]. The output

of this step is the standardized WSI with the staining characteristics similar to the predefined template WSI.

3.3 Empirical evaluation

3.3.1 Histopathology Image dataset

Two histopathological image datasets were used for empirical evaluation of the proposed algorithm. The first dataset consisted of 125 digitized H&E stained WSIs of lymph nodes from 3 patients. These slides were serially sectioned and stained in 5 different Dutch pathology laboratories, each using their own routine staining protocols. The set up included staining of slides on different days of the week. Staining protocol variations between laboratories include temperature, concentration, staining time and manufacturer of different solutions. Frequency of refreshing staining solutions may also differ between laboratories. All slides were digitized using a CCD RGB camera (Zeiss AxioCam HRC) mounted on a light microscope (Zeiss AxioPlan 2im) with a $40\times$ objective lens. Each image has square pixels of size $0.256\mu m$ $\times$ $0.256\mu m$ in the microscope image plane.

The second image dataset comprised three batches, each containing 10 H&E stained histological slides of rat liver with different amounts of confluent necrosis. The slides were stained in University Hospital Jena, Germany, and scanned with a Hamamatsu Nanozoomer Scanner at $40\times$ objective magnification. A human expert generated ground truth data on this dataset by annotating necrotic tissue in each section at small tile levels of size 256×256 pixels in the highest resolution.

3.3.2 Experiments and Results

To evaluate the performance of the WSICS algorithm three experiments were performed:

1. A quantitative, comparative evaluation of the performance of the proposed algorithm versus two competing algorithms in calculating stain vectors

2. A qualitative and quantitative evaluation of the performance of the proposed standardization algorithm in comparison with other methods in achieving color constancy

3. A quantitative evaluation of the effect of employing the proposed whole-slide standardization algorithm on the performance of a necrosis quantifying CAD system[61]

Experiment 1

The aim of this experiment was to quantitatively evaluate the performance of each standardization algorithm in extracting stain vectors using a subset of the lymph node dataset. For the experiment, we randomly selected 5 different slides for each laboratory and took one sample FOV image from each slide. This yielded a total of 25 images from 25 slides. For each of the images, a large number of pixels were manually annotated to give representative ground truth pixels for hematoxylin and eosin classes. Subsequently, hematoxylin and eosin stain vectors were calculated using Ruifrok's color deconvolution method[42].

The performance of our proposed algorithm is compared to that of two state-of-the-art algorithms for stain deconvolution: the appearance normalization algorithm by Macenko at al.[43], and the nonlinear mapping approach to stain normalization by Khan et al.[56]. The algorithm by Macenko at al.[43] tries to find the fringe of pixel distribution in the optical density space to determine the stain vectors. The algorithm by Khan et al. utilizes a pretrained relevance vector machine (RVM) classifier to classify the pixels in the image into different stain components in the image. The stain vectors are then calculated from the set of labeled pixels for each stain class. In this regard, this method works similar to our proposed algorithm, however, our algorithm does not require manual training of the classifier. The other fundamental difference of our method is that it mainly uses shape information for identifying hematoxylin pixels, and samples from pixels outside of ellipse-shaped objects for eosin.

Figure 3.4 shows the results of classifying pixels as hematoxylin, eosin, or background, which are intermediate steps in our color standardization algorithm and that of Khan et al.[56]. As seen in this figure, the algorithm by Khan performs poorly on the third and fourth example which can be related to the fact that the staining colors in these images deviate from the image batches that were originally used for training the RVM classifier. Contrary to this approach, our fully automatic algorithm effectively derives the hematoxylin and eosin distributions and classifies the pixels very accurately mainly due to using shape information.

The mean and the standard deviation (SD) of the Euclidean distances between the stain vectors from the annotated data and the stain vectors derived by each of the algorithms is used as a measure to compare the efficacy and robustness of the algorithms in extracting the correct stain vectors. Table 3.1 presents the results. Our proposed algorithm is performing considerably better in calculating stain vectors for both hematoxylin and eosin stains.

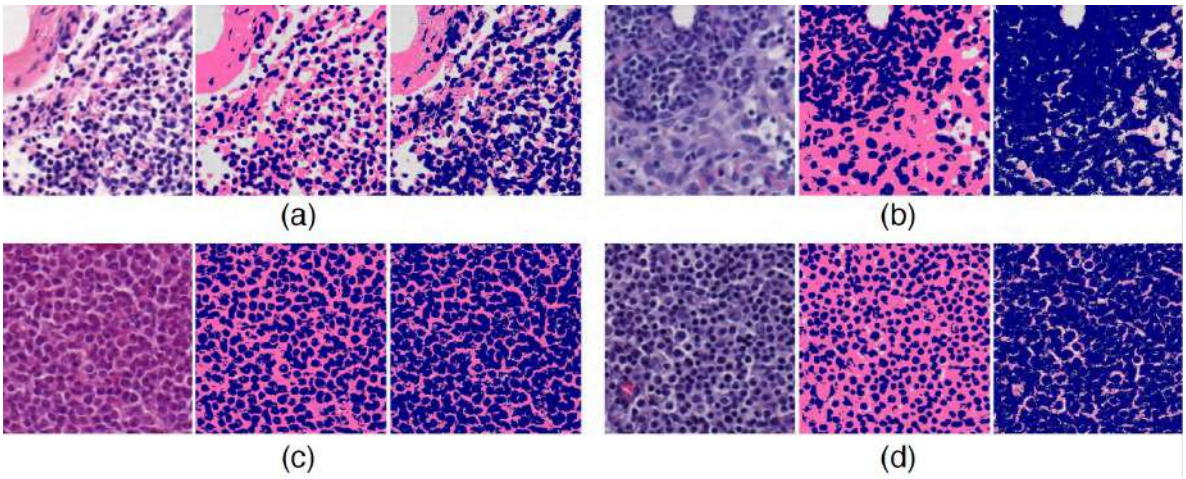

Figure 3.4: Comparison of the pixel classification performances between the method by Khan et al.[56] and our proposed algorithm. (a) - (d) show 4 different classification results produced by the two algorithms. For each sub-figure the image on the left is the original image to be classified. The middle image presents the classification result by our algorithm and on the right is the result produced by the algorithm of Khan et al.[56].

Table 3.1: Average Euclidean distances between the manually determined stain vectors in the original images and the stain vectors computed by different algorithms. d_H and d_E denote the average Euclidean distances for the hematoxylin and eosin vectors, respectively.

Method	d_H (mean $\pm$ SD)	d_E (mean $\pm$ SD)
Mackenco[43]	0.2120 ± 0.0314	0.1764 ± 0.0296
Khan[56]	0.0393 ± 0.0134	0.1187 ± 0.0297
Proposed	$\mathbf{0.0135 \pm 0.0093}$	$\mathbf{0.0186 \pm 0.0225}$

Experiment 2

The aim of this experiment was to qualitatively and quantitatively evaluate the performance of the WSICS algorithm. We focus on inter-laboratory variations of the H&E staining in the lymph node dataset, as this is a major concern in large scale application of CAD in pathology. The performance of our proposed algorithm is compared to that of three previously published algorithms: global standardization (GS) by bagci et al.[44], the appearance normalization algorithm by Macenko et al.[43], and the nonlinear mapping approach to stain normalization by Khan et al.[56]. Five representative field-of-view (FOV) images were acquired from each WSI yielding a total of 625 images. Each image is of size 1388×1040 pixels. The results of the standardization performed by different algorithms are shown in Figure 3.5. The image shown in the top left of Figure 3.5.a was used as the template image to extract

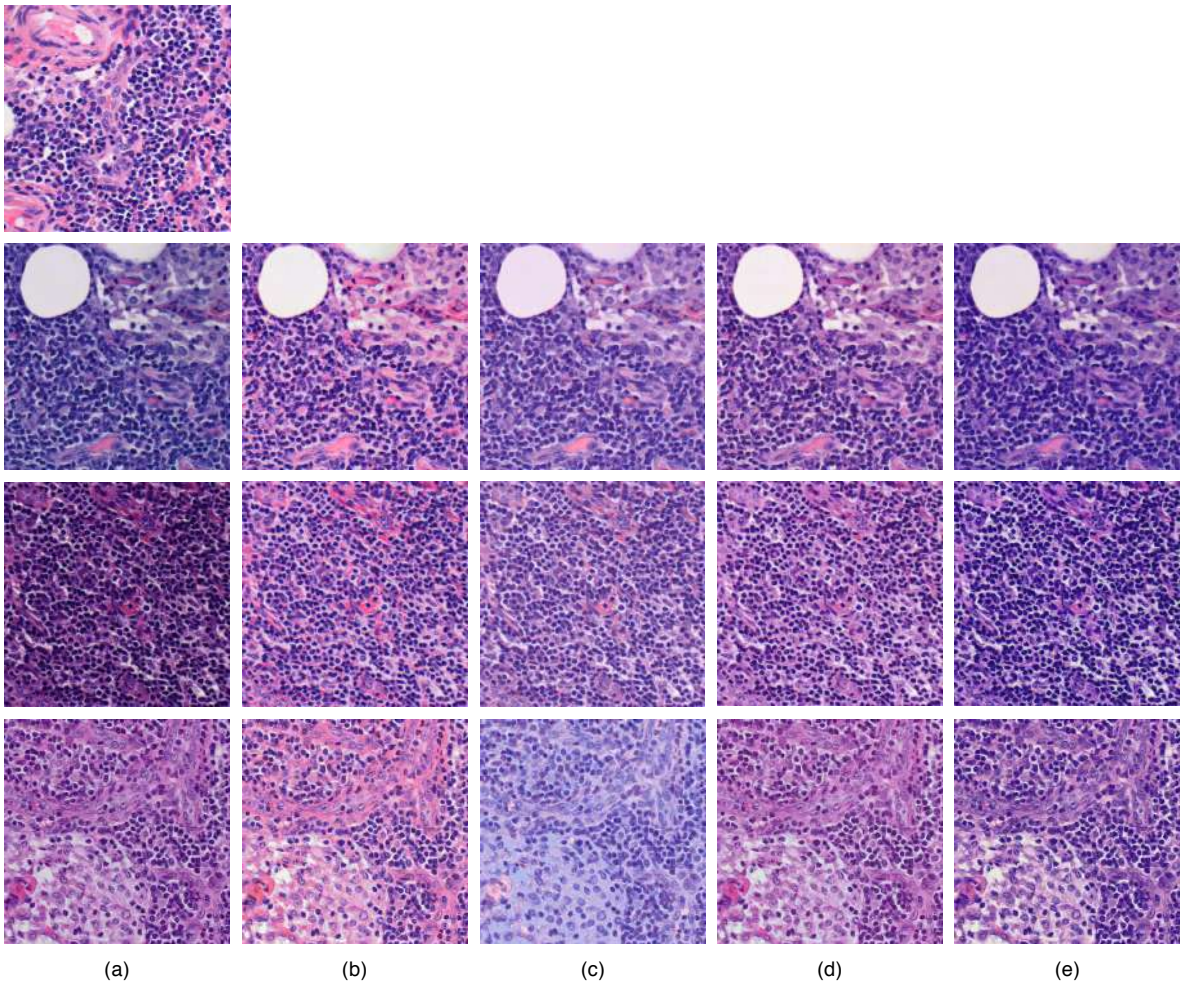

Figure 3.5: Illustration of the performance of different stain standardization algorithms. The top left image has been used as the template image. Column (a) three original images sampled from different slides stained in different laboratories. (b) the result of standardization using the WSICS, (c) the algorithm devised by Macenko et al.[43], (d) Bagci et al.[44], and (e) the algorithm by Khan et al.[56].

Table 3.2: Standard deviation and coefficient of variation of NMI for all the images in the five laboratories.

	Laboratory 1		Laboratory 2		Laboratory 3		Laboratory 4		Laboratory 5	
Method	NMI SD	NMI CV	NMI SD	NMI CV	NMI SD	NMI CV	NMI SD	NMI CV	NMI SD	NMI CV
Original	0.0206	0.0405	0.0254	0.0485	0.0305	0.0641	0.0279	0.0466	0.0201	0.0348
Bagci[44]	0.0157	0.0294	0.0184	0.0336	0.0163	0.0307	0.0256	0.0422	0.0172	0.0303
Macenko[43]	0.0180	0.0285	0.0161	0.0256	0.0092	0.0148	0.0203	0.0317	0.0127	0.0200
Khan[56]	0.0154	0.0313	0.0194	0.0385	0.0199	0.0390	0.0214	0.0398	0.0139	0.0254
WSICS	**0.0083**	**0.0147**	**0.0075**	**0.0131**	**0.0081**	**0.0141**	**0.0087**	**0.0155**	**0.0052**	**0.0092**

parameters required by different standardization algorithms. The three standardized images by the method proposed by Macenko et al.[43] have more color variability compared to the other methods. Moreover, it can be seen that the color of the images standardized by the WSICS algorithm have the highest similarity to the color of the template image as compared to the other approaches.

In this experiment, the choice of the template slide was based on the opinion of two pathologists, who studied a large number of slides from each laboratory. The major criteria for them to select a high quality staining are: (1) high contrast between hematoxylin and eosin staining (2) visibility of the nuclear texture. The majority of the slides stained in Lab 1 were found to meet these criteria and therefore a slide from this batch was selected as the template slide in this experiment.

Quantitative analysis of standardization results is based on color constancy of nuclear staining and eosin staining independently. To evaluate the color constancy of the nuclear staining, nuclei were first detected using fast radial symmetry transform[65]. The detected candidate nuclei were subsequently segmented using a marker-controlled watershed algorithm as illustrated in[66]. Quantitative measures of the area and elliptical shape were computed for each candidate nucleus. The elliptic variance descriptor (E_{var})[67] was used to measure how closely the borders of a fitted ellipse agree with those of the segmented nucleus-like object. Objects that were too small (area < 200) or irregular ($E_{var} > 0.13$) were rejected as artifacts. The normalized median intensity (NMI) measure[45,50] was then chosen to evaluate color constancy of the nuclei. This measure enables comparison of the intensity statistics over a population of images. The NMI measure is defined as:

$$NMI\left(I\right) = \frac{\underset{i \in I}{Median}\left\{U\left(i\right)\right\}}{\underset{i \in I}{P_{95}}\left\{U\left(i\right)\right\}} \tag{3.5}$$

where $U\left(i\right)$ denotes the average of the R, G, B values for the pixel i in image I, and P_{95} denotes the 95th percentile. Note that to increase the robustness of the NMI measure against noisy pixels in the image, instead of dividing the median term by the maximum, we divide it by the 95th percentile. The standard deviation of the NMI values (SDN) and coefficient of the variation (i.e. standard deviation divided by mean) of the NMI values (CVN) were computed for the images in different laboratories before and after standardization using different methods. The results are shown in Table 3.2. Note that small values for SDN and CVN indicate that nuclei in images from different laboratories have similar color distributions (i.e. qualitatively look the same) after stain standardization. In all cases, the WSICS algorithm yielded the smallest SDN and CVN. The box plots of the NMI values for each laboratory is shown in Figure 3.6. The box plot shows that the spread of NMI values about the

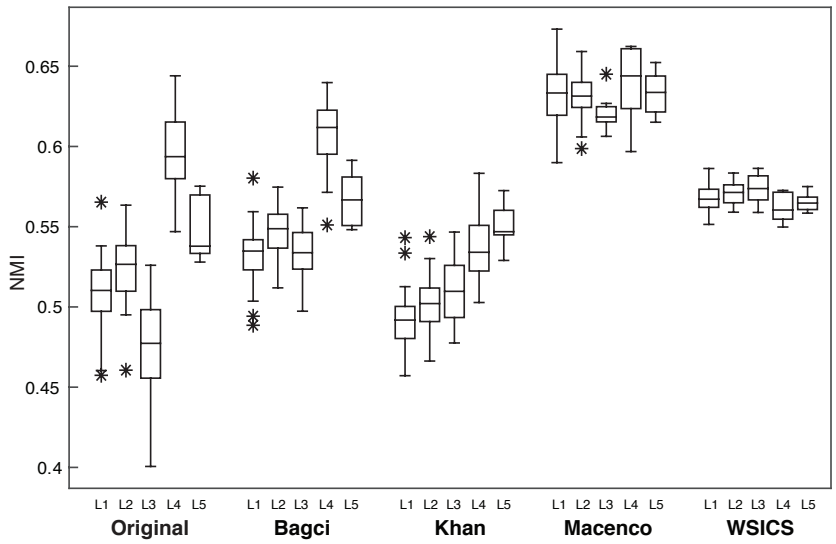

Figure 3.6: Box plots of the normalized median intensity values of the slides stained in five laboratories for all the methods in Experiment 3.

median (inter-quartile range) is the smallest for the proposed algorithm compared to the competing state-of-the-art algorithms. In addition, the distribution of NMI values across different laboratories is considerably more stable for the WSICS algorithm.

To evaluate the color constancy of the eosin staining we used part of the lymph node dataset that was used in experiment 1. Provided that automatic segmentation of eosin stained tissue structures is more complicated, we evaluate the color constancy of eosin staining within regions that were manually annotated as such. The results are shown in Table 3.3. Note that the SDN and CVN are computed over all the 25 images from the 5 laboratories. Overall, WSICS yielded the smallest SDN and CVN.

Table 3.3: Standard deviation and coefficient of variation of NMI for the eosin dye for the 25 images considered.

Method	NMI SD	NMI CV
Original	0.0563	0.0748
Bagci[44]	0.0200	0.0247
Macenko[43]	0.0362	0.0439
Khan[56]	0.0434	0.0555
WSICS	**0.0191**	**0.0220**

The computation time for each algorithm to standardize an image of size 1388×1040 (averaged over 20 images) is presented in Table 3.4. The computation time required by WSICS to create a look-up table for standardizing a WSI of lymph node

is 10 ± 1 minutes.

Table 3.4: Computation time (in seconds) for presented methods for standardizing an image of size 1388×1040.

Method	Mackenco [43]	Khan [56]	Bagci [44]	WSICS
Processing time	1.89	123.73	0.87	18.94

Experiment 3

In this experiment, the performance of an already published CAD system [61] for quantifying necrosis was evaluated before and after standardization of the slides. A dataset comprising three batches of H&E stained histological WSIs of rat liver sections with different amounts of confluent necrosis was available. The CAD system described in [61] using a Random Forest classifier was used to detect necrotic tissue in WSI. This system utilizes local binary pattern (LBP) [68] and pixel value statistics features for each of the individual channels of the RGB and HSV color models. The performance of this system was assessed using a leave-one-batch-out cross-validation scheme. At each cross-validation round, CAD was trained with all slides from two of the batches and validated on the third batch. The same assessment was carried out after standardizing the entire slides using one of the slides in the training set as template image. The performance of the CAD system was then assessed for each cross-validation round in terms of the area under the receiver operating characteristic (ROC) curve [69] at the patch level of size 256×256. Figure 3.7 shows the ROC curves for each cross-validation round. The area under ROC curves with and without employing WSICS were compared using the bootstrap test [70]. This test was used to test the null hypothesis that the CAD system performs equally well with and without standardization, versus the alternative hypothesis that it does not. The AUC results for each cross-validation round and the corresponding p-values are summarized in Table 3.5. The p-value for the test was smaller than 0.01 for all the 3 ROC curve pairs, thus providing evidence that CAD performance is increased by applying the WSICS algorithm.

3.4 Discussion and conclusion

In this paper, we presented a novel algorithm, called whole-slide image color standardizer (WSICS), for standardization of whole-slide histopathological images. We showed that the WSICS algorithm outperforms previously published algorithms.

Table 3.5: Summary of the AUC results for each cross-validation round of experiment 3 and the corresponding p-values of the bootstrap test to compare AUCs

Train set	Test set	AUC	AUC with WSICS	p-value
1 and 2	3	0.939	0.963	< 0.01
1 and 3	2	0.930	0.985	< 0.01
2 and 3	1	0.310	0.944	< 0.01

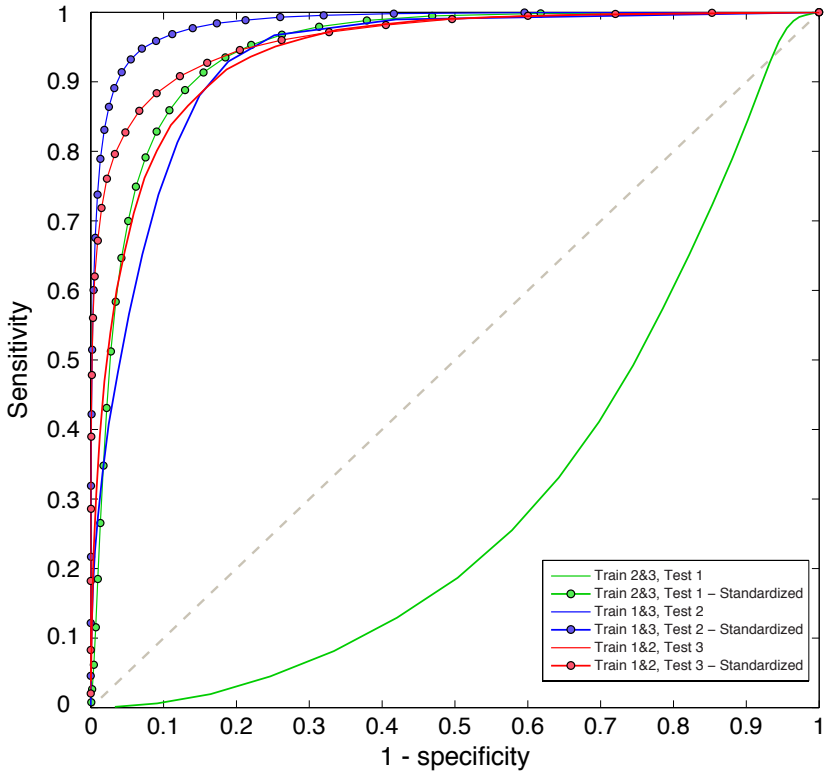

Figure 3.7: Performance of the CAD system with and without standardization for different cross-validation rounds

Even more importantly, we showed that the algorithm was capable of significantly improving the quality of an existing CAD system, rendering it applicable even for specimens exhibiting staining characteristics which strongly deviated from the specimens it was trained on.

The main characteristics of our algorithm are: 1) fully automated detection of stain components in WSIs enabling unsupervised operation, 2) robust stain classification by making use of spatial information, 3) the HSD color model for transformation of chromatic and density distributions, and 4) application of weights to create a smooth and artifact-free whole-slide standardization. The algorithm was shown to be very robust against all sources of staining variation.

Our algorithm for classification of tissue components avoids manual training of

the classifier by learning samples of each class from the image at hand using prior shape information. This yields fully automated, objective and reproducible classification results on image data with various sources of variation. Unlike the algorithms in the literature which rely solely on color information to identify stain components our algorithm incorporates spatial information which makes it significantly more robust. Figure 3.4 compares the classification result obtained by our proposed algorithm and the algorithm by Khan et al. Our approach performed remarkably well on the entire image data provided by 5 different (academic and non-academic) laboratories containing severe staining variation. We observed no classification failure over the entire dataset. In the classification approach in[55], in contrast, the automatically extracted reference stain vectors had to be replaced in more than 10% of the images due to segmentation failures. Our previous experience with unsupervised approaches such as EM-based segmentation[45,58] shows that segmentation may occasionally fail. This failure is mainly due to the cases in which the chromatic distribution of the channels overlap significantly. The algorithm proposed by Macenko et al.[43], moreover, becomes unstable in images with poor contrast and insufficient data for each stain.

Existing color transformation approaches align different channels of a particular color model, independently, using separate $1D$ transformations[45 55 57]. These approaches, however, assume that the channels of the color model utilized are independent which generally does not hold. The use of a more advanced color model called HSD which was specifically designed for transmission light microscopy enables independent transformation of chromatic and density information. We performed non-rigid registration of $2D$ chromatic distribution using several transformations which provides better aligning of the color information compared to separate $1D$ transformation of each channel. The use of class-dependent weights in combining these transformations yields a smooth standardization result. Consequently, our method consistently shows improved color constancy compared to existing methods.

We presented 3 experiments to evaluate the efficacy of the WSICS method. In the first experiment, we presented a quantitative, comparative evaluation of the performance of the proposed algorithm versus two state-of-the-art algorithms in extracting stain vectors by Macenka et al.[43] and Khan et al.[56]. The results show that for both hematoxylin and eosin dyes, the derived stain vectors by the WSICS algorithm are substantially more accurate and highly comparable with the stain vectors computed from manually annotated regions in the image.

In the second experiment we presented qualitative and quantitative evaluations of our method relative to three state-of-the-art methods: global standardization (GS)

by Bagci et al.[44], the appearance normalization by Macenko et al.[43], and the nonlinear mapping approach by Khan et al.[56]. Qualitative assessment of the results show the efficacy of our algorithm in enhanced color constancy of the histology images.

The results shown in Figure 3.5 demonstrate that the images with severe staining variation can be standardized to resemble the template image using the WSICS algorithm. Compared to the three state-of-the-art methods, our method performs considerably better in standardizing eosin staining which is in correspondence with the results achieved in experiment 1. The algorithm devised by Macenko[43] yields poor result with artifacts in the third example which is due to wrong estimation of the stain vectors. This algorithm tries to find the fringe of pixel distribution in the optical density space to determine the stain vectors. As seen in the third example shown in Figure 3.5.a there is a significant difference in the staining of blood cells compared to cytoplasmic/stromal staining. The poor contrast between nuclear staining and cytoplasmic/stromal staining has led to poor standardization results by this method. The Global Standardization algorithm, on the other hand, achieves smooth standardization output without any artifacts (see Figure 3.5.d). However, the algorithm is clearly unable to match the staining quality to the template image. This is mainly due to the usage of a single transformation function for standardizing the image which does not correspond with the existence of multiple components in the data. The algorithm proposed by Khan et al.[56] yields artifact-free standardized images. However, the quality of the standardized images by this method significantly deviate from the template image. This algorithm uses an RVM classifier to classify the pixels into different stain components in the image. Using a pretrained classifier makes this algorithm unstable in case where the color of the dyes in the test specimen deviate from the image batches that the RVM classifier was initially trained on. As a result, the estimation of stain vectors may fail. This was also observed in experiment 1. As shown in Figure 3.4, heavy pollution of the hematoxylin population with pixels from connective tissue and cytoplasm leads to wrong estimation of the stain vectors by this algorithm. As a result, all the eosin stained structures will have a purplish-blue color after standardization. Contrary to this approach, our algorithm effectively defines the hematoxylin and eosin distributions by incorporating spatial information. This, in turn, leads to better standardization performance by our proposed algorithm.

The results of the quantitative assessment, as summarized in Table 3.2 and Table 3.3, show that for the entire lymph node dataset stained in 5 different laboratories, the algorithm proposed here outperformed the competing standardization methods by yielding the lowest standard deviation and coefficient of variation of the NMI measure.

The application of whole-slide standardization to computer-aided diagnosis of histopathology data has so far remained elusive in the literature due to technical complexities in dealing with whole-slide images. The focus has been limited to investigating the contribution of standardization algorithms that work at the small image patch level. In the third experiment, we evaluated the impact of our proposed whole-slide standardization algorithm on the performance of a necrosis quantification CAD system. The performance of the CAD system for quantifying necrosis was assessed in a leave-one-batch-out cross validation experiment, before and after standardization. The performance of the CAD system was better in all cross-validation rounds after utilizing the WSICS algorithm. In particular, the performance was significantly improved for the case that the CAD system was trained on batch 2 and 3 and tested on the first batch. The ROC curve shown in Figure 3.7, illustrates that the performance without standardization is worse than random guessing. The reason for this can be related to the significant difference in the staining color of the slides in batch 1 relative to other batches which results in the viable tissue having intensity and color ranges similar to necrotic tissue in other batches, and vice-versa. This is, however, effectively addressed using the proposed method. Hence, the substantial impact of employing the proposed whole-slide standardization algorithm on the performance of an already published CAD system further demonstrates its efficacy and reliability.

One limitation of the current study is that although the WSICS algorithm has been designed to standardize whole-slide images, the comparison of the color constancy of the images standardized by different algorithms in experiment 2 is limited to image patches (sub-images from the WSI). This is because the competing algorithms have been designed to standardize patch images only.

The WSICS algorithm has been specifically described for standardization of H&E stained images. The algorithm, however, can be adapted to work with other histological staining techniques such as immunohistochemistry (IHC). Identification of the stain components is more straight forward in IHC. The major reason is that in IHC, in contrast to H&E staining, the chromatic distribution of the stain components have a small overlap in the chromatic plain of the HSD model. In future work, we will concentrate on adapting the proposed algorithm to be utilized on other staining techniques and investigate the possibility of extending our method for standardizing WSIs with more than two stain components.

Acknowledgment

The authors wish to acknowledge the financial support by the European Union FP7 funded VPH-PRISM project under grant agreement n°601040. We also gratefully acknowledge financial support from the Stichting IT Projecten Nijmegen (NT) and the Maurits en Anna de Kock foundation for image analysis equipment. The authors also wish to acknowledge support from the histology laboratory of the Department of Pathology, Radboud University Medical Center, Nijmegen, the Netherlands and Prof. Dr. med. Uta Dahmen and Dr. med. Olaf Dirsch, University Hospital Jena, Jena, Germany.

Detection of regions of interest in whole-slide histopathological images

4

Babak Ehteshami Bejnordi, Geert Litjens, Meyke Hermsen, Nico Karssemeijer and Jeroen AWM van der Laak

Original title: A multi-scale superpixel classification approach to the detection of regions of interest in whole-slide histopathology images

Published in: Proceedings of SPIE Medical Imaging, pp. 94200H-94200H, 2015.

Abstract

This paper presents a new algorithm for automatic detection of regions of interest in whole-slide histopathological images. The proposed algorithm generates and classifies superpixels at multiple resolutions to detect regions of interest. The algorithm emulates the way the pathologist examines the whole-slide histopathology image by processing the image at low magnifications and performing more sophisticated analysis only on areas requiring more detailed information. However, instead of the traditional usage of fixed sized rectangular patches for the identification of relevant areas, we use superpixels as the visual primitives to detect regions of interest. Rectangular patches can span multiple distinct structures, thus degrade the classification performance. The proposed multi-scale superpixel classification approach yields superior performance for the identification of the regions of interest. For the evaluation, a set of 10 whole-slide histopathology images of breast tissue were used. Empirical evaluation of the performance of our proposed algorithm relative to expert manual annotations shows that the algorithm achieves an area under the Receiver operating characteristic (ROC) curve of 0.958, demonstrating its efficacy for the detection of regions of interest.

4.1 Introduction

Automated detection of clinically meaningful Regions of Interest (ROIs) in whole-slide histopathological images is an important initial step in the development of an automated computer-aided diagnosis system. Accurate extraction of these ROIs would allow performing complex image analysis tasks only on specific relevant areas within the whole-slide image (WSI). This is in particular of utmost importance for an efficient analysis of large histopathological images. Two major approaches have been utilized in the literature for the development of automated CAD systems to detect cancer in whole-slide histopathological images. The first is to perform image analysis operations at a single specific image resolution to classify different tissue structures[71,72]. The second utilizes a multi-resolution scheme to classify high-resolution WSI[73–76]. Contrary to the first approach which does not correspond to the multi-scale approach used by the pathologists, the second approach emulates the way pathologist examines a histology slide. The multi-resolution approach significantly reduces the computational time required to analyze the whole-slide by processing the image tiles at low magnifications with the least computational burden and performing more sophisticated analysis of the corresponding tiles in higher magnification only when the decision for the classification requires more detailed information[76]. To achieve this, these algorithms make use of small fixed-size rectangular patches and try to classify them into different tissue classes. Fixed sized patches, however, can span multiple distinct tissue structures, thus degrading the classification performance.

Superpixels are alternative visual primitives which can compensate for the shortcomings of pixels and patches. A superpixel is a perceptually meaningful atomic region that aggregates visually homogeneous pixels while respecting object boundaries. Superpixels are obtained from the over-segmentation of the image. As boundary information is respected during partitioning the image into superpixels, more accurate segmentation results can be obtained by allocating superpixels to the appropriate target class. Superpixels have been increasingly used in medical imaging applications and greatly reduce the complexity of image processing tasks. Superpixel classification approaches have been successfully applied in several applications such as segmentation of brain MRI images[77], and prostate cancer detection and classification[78].

In this paper, we propose a multi-resolution superpixel classification approach to detect ROIs in whole-slide histopathology images. The proposed system, initially partitions the image, in the lowest magnification, into a set of non-overlapping superpixels. At the lowest magnification, superpixels are classified into regions con-

taining tissue and regions belonging to the background. New superpixels are constructed at the intermediate magnification within the areas containing tissue and are classified into a particular tissue component (e.g. stroma, background, epithelial nuclei). Finally, a new set of high-resolution superpixels is built at the highest magnification only in the areas where the classifier, at the lower level, yielded a low confidence in assigning the output label. A second stage classifier is then employed to classify those superpixels more accurately. We present an empirical evaluation of the performance of our algorithm on H&E stained WSIs of breast tissue and present a comparison with the traditional tile analysis algorithm for finding ROIs.

4.2 Methods

Our algorithm for the detection of regions of interest has three main components. The first is identification of areas containing tissue by classifying superpixels built on the lowest magnification. The second constructs new superpixels at the intermediate magnification on the areas containing tissue and classifies them into different tissue components. The third classifies newly built superpixels at the highest magnification for the regions requiring more detailed information for accurate classification. Detailed description of different steps of the proposed algorithm are discussed below.

4.2.1 Tissue identification in low magnification

In this paper, the Simple linear iterative clustering (SLIC)[79] algorithm was used to generate superpixels. SLIC algorithm offers strong performance in terms of adherence to edges and segmentation speed, hence very well suited to histopathological image analysis. The proposed implementation of the SLIC algorithm performs image clustering in the CIELAB color space. However, we performed a transformation into the hue-saturation-density (HSD) color model, which was specifically designed for absorption light microscopy[46]. The HSD model transforms RGB data into two chromatic components (c_x and c_y; which are independent of the amount of stain) and a density component (D; linearly related to the amount of stain). Tissue identification is achieved by first partitioning the image into superpixels at the lowest magnification. This is followed by a classification phase to distinguish between foreground objects (tissue) and background. A pixel inside a superpixel is classified into the background class if its overall density is lower than 0.2 and the density of its r, g, and b channels is lower than 0.25. Superpixels containing more than 90% of background pixels are classified as background objects. At the end of this stage, a

whole-slide mask is generated for areas comprising of tissue.

4.2.2 Tissue component classification at the intermediate magnification

For computer-aided diagnosis of breast cancer the epithelial regions of the tissue are of major clinical importance[25]. Although the importance of the stromal features for prognosis of breast cancer has been recognized[80], focusing on the detection of epithelial regions does not limit the applicability of such features. Stromal features can still be computed from the stromal areas surrounding the suspicious epithelial tissue. Therefore, automated diagnosis of cancer requires identification of epithelial regions as its initial step. For this reason, the tissue was classified into three components: epithelium, stroma, and background. To classify the entire WSI into these tissue components, superpixels were generated at the intermediate magnification over the entire regions which contain tissue. In practice, computational resource requirements do not allow to generate superpixels on the entire WSI at once. Therefore we need to generate superpixels separately on small image tiles containing tissue. However, this will lead to undesirable superpixel structures at the borders of the image. Figure 4.1 illustrates the result of generating superpixels on two consecutive tiles. As can be seen, the shape of the superpixels in the transition area between the two tiles is affected by the tile boundary. In the following section, we illustrate our proposed method for generating continuous superpixels over the entire WSI.

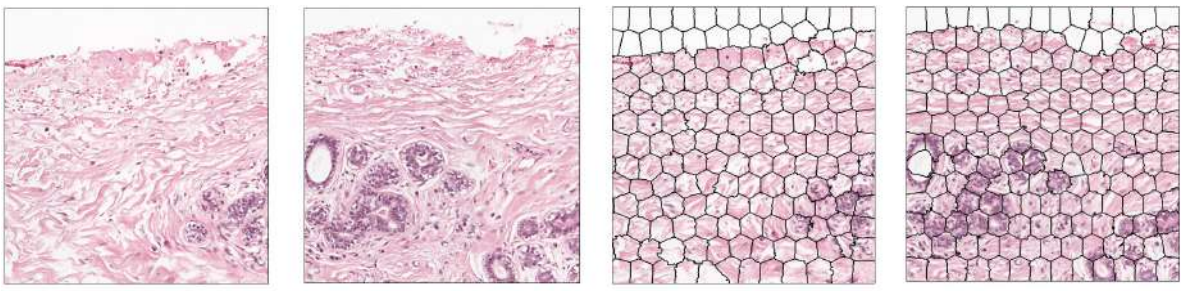

Figure 4.1: Illustration of generating superpixels on consecutive image tiles. (a), (b) Original images of the first and second tile. (c), (d) the output of superpixel generation on *(a)*, and *(b)*.

Generation of continuous superpixels over the WSI

To address the problem associated with undesirable superpixel boundaries at the edge of each tile, superpixels are generated on overlapping tiles. The size of the over-

lap area is determined in such a way that covers at least 2 layers of superpixels from the previous tile. Figure 4.2 illustrates how the generation of continuous superpixels on overlapping tiles is performed. First, the original image shown in Figure 4.1a is extended by the addition of the overlapping area from the next tile. Superpixels are then generated on this image yielding the image shown in Figure 4.2a. To build superpixels on the next overlapping tile, we replace the overlapping area (on the left side) of the second tile image using the mask shown in Figure 4.2b. This mask is extracted from the overlapping area from Figure 4.2a, in which the values of the superpixels attached to the image boundary (on the right side) are set to one and the rest to zero. By multiplying this mask with the corresponding overlapping area of the second tile image and building new superpixels on the image the result in Figure 4.2c is obtained. As shown in this Figure, the black area creates a strong transition of pixel intensity values in this image which will consequently force the superpixels to adhere to the strong artificially created boundaries hence yielding a continuous superpixel arrangement in the transition area of the two tiles. The final result after stitching the tiles is presented in Figure 4.2d. In practice, the same technique is applied to the other sides of each patch, to preserve the superpixel continuity from all sides.

Superpixel classification at the intermediate magnification

In the next stage, a classifier is constructed which operates on the regions defined by the superpixels at the intermediate magnification. A total of 54 features were extracted for the classification task including local binary patterns and statistics derived from the histogram of the three channels of the HSD color model. Training data was acquired from a set of superpixels which were annotated as epithelium, stroma, and background. Identifying superpixels belonging to the background class was done by setting a threshold on the median density of the superpixels. The remaining superpixels were classified as epithelium or stroma using a random forest classifier. To defy the curse of dimensionality and to reduce the feature computation time, a feature selection experiment was carried out. Two feature selection methods were utilized: multiple support vector machines with recursive feature elimination (MSVM-RFE)[81] and guided regularized random forest (GRRF)[82]. Feature selection by these two methods was achieved by 100 iterations of 5-fold cross validation, each iteration with random combinations of samples in the training and the test set. The output from both methods showed that Local binary patterns of the three channels do not contribute significantly to the classification accuracy, hence excluded from the classification task in the intermediate level. The random forest classifier was therefore built on the training data using the selected features. Non-background su-

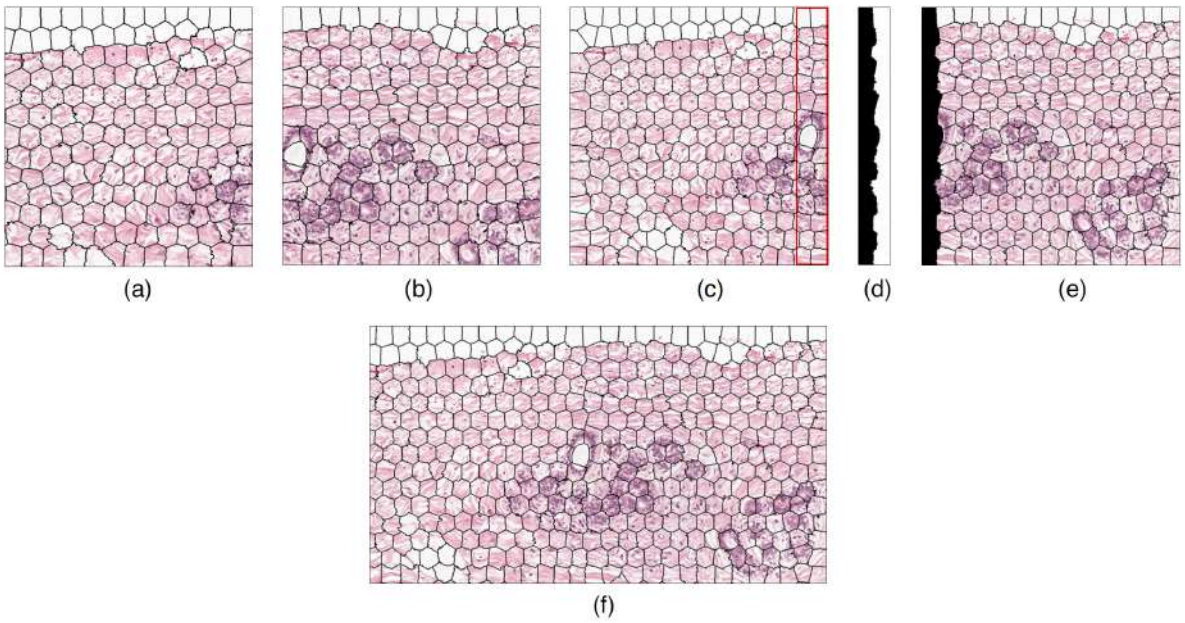

Figure 4.2: Illustration of generating continuous superpixels on overlapping tiles. (a) and (b) show the results for building superpixels on the image patches shown in Figure 4.1a and Figure 4.1b. (c) Generating superpixels on Figure 4.1a by the inclusion of the overlapping area (rectangle in red) from the next tile. (b) The mask extracted from the overlapping area in *(a)* is achieved by setting the value of the superpixels attached to the right border of the image to 1 and the rest of pixels to 0. (c) Building superpixels on the next overlapping tile. The overlapping area of the tile is replaced using the mask generated in *(b)* before generating the superpixels. (d) The result of stitching the tiles yielding continuous superpixels over the entire WSI. Note that the last layer of the superpixels attached to the right side of the image in *(a)* and the first layer of superpixels attached to the left side of the image in *(c)* are removed for stitching the two tiles.

perpixels were classified using the trained model. Based on the confidence of the classifier for assigning a label to the superpixel, we decide if more detailed information is needed to classify the region. If the probability of the superpixel belonging to a specific tissue class exceeds 90%, no further analysis is required. We perform more detailed analysis not only when the classifier has a low confidence (lower than 90%) but also when the classifier assigns the epithelium label to a superpixel. The reason for this is that we want to classify epithelium regions with more accurate boundaries which are often achievable at higher magnifications.

4.2.3 Tissue component classification at the highest magnification

A second stage classifier was constructed to classify only the areas which were marked as requiring more detailed analysis. For this purpose, a new set of superpixels were generated at the highest magnification on these areas. Figure 4.3b shows how the new set of superpixels are generated in areas requiring more detailed analysis. The newly built superpixels were classified into epithelium, stroma, and background class with the same approach illustrated in lower magnification using a second random forest classifier trained on superpixels annotated in higher magnification. A similar feature selection experiment was carried out for the classifier at this magnification. Unlike the intermediate level classification problem, local binary patterns had discriminatory power for the classification. All of the 54 extracted features were therefore used for the second random forest classifier. Finally, we performed a post-processing for the superpixels which were classified with a low confidence on the highest magnification. The new probability for these superpixels was calculated using the average probabilities of their neighboring superpixels.

4.3 Empirical evaluation

4.3.1 Histology images

The image data used in this study originate from a set of 10 digitized H&E stained histopathology slides of Breast tissue sampled from 10 patients. Each slide was reviewed by a pathologist and assigned a pathological diagnosis. The dataset contains two samples from each of the following cases: Normal, Ductal carcinoma in situ, Invasive ductal carcinoma, Lobular carcinoma in situ, and Invasive lobular carcinoma. The whole-slide histopathology images were acquired using 3DHistech Pannoramic 250 Flash II scanner at $20X$ magnification. To generate ground truth data for evaluating the performance of the algorithm, two trained subjects were recruited to delineate epithelial regions within the entire slide using ImageScope viewer tool.

4.3.2 Experiments and results

To evaluate the performance of our proposed algorithm, a comparison was made against the traditional tile analysis. The exact same scheme was employed to identify ROIs by this method. The ability of the SLIC algorithm to generate approximately equal sized superpixels enables us to make a fair comparison with the tile analysis method. Consequently, each tile image was divided into rectangular arranged square patches which have the same size as the average superpixel size in SLIC al-

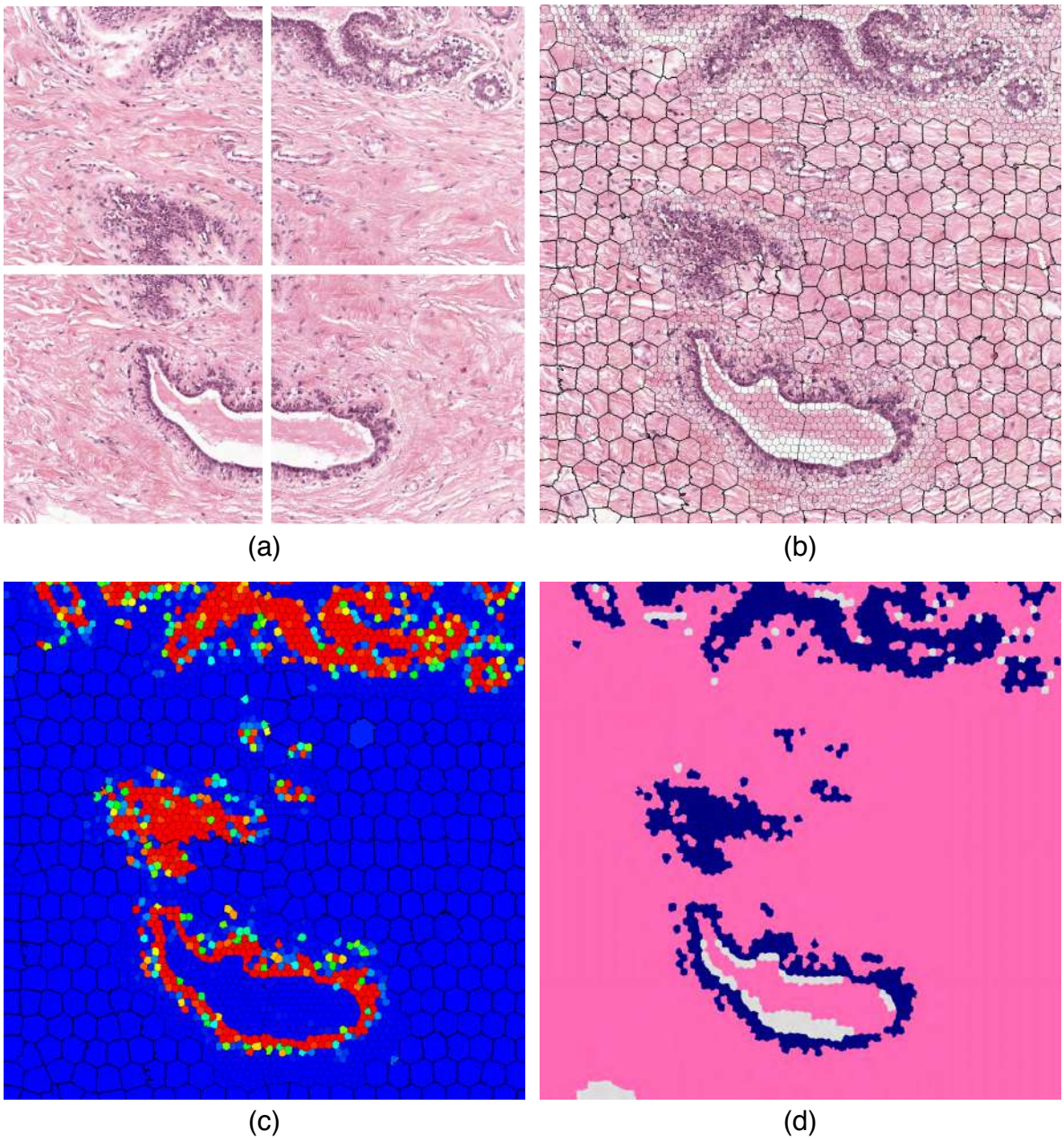

Figure 4.3: Illustration of multi-scale superpixel classification. (a) The original tiles. (b) Generation of superpixels in multiple levels. (c) Likelihood map showing the probability of a superpixel to belong to the epithelium class. Larger superpixels were classified with high confidence in low magnification. Smaller superpixels are generated at highest magnification and classified using the second stage classifier. (d) The result of hard classification by incorporating context information for superpixel having low confidence in their classification output.

gorithm. Tiles were classified at different magnification using the same classifiers used by the multi-scale superpixel classification algorithm.

The performance of the two systems was evaluated in terms of the area under

the receiver operating characteristic (ROC) curve. Figure 4.3 illustrates the selected steps for the classification process with our proposed algorithm for 4 neighboring tiles stitched together and the likelihood map of the classifier probability output. The superpixels classified as the background class have been excluded from the analysis because the classification task for this class is very simple and including them might result in an optimistic measure for the false positive rate. The area under the ROC curve (AUC) reflecting the overall performance of the multi-scale superpixel classification algorithm was 0.958. The AUC for the tile analysis in comparison was 0.932

4.4 Discussion and conclusion

This paper presented a novel multi-scale superpixel classification approach to detect regions of interest relevant to the diagnosis of breast cancer in whole-slide histopathology images. The multi-resolution whole-slide analysis allows identification of areas easy to classify in low magnifications and classifications of areas requiring more detailed analysis in higher magnifications. This approach significantly reduces the computational time required to analyze the whole-slide compared to pixel classification methods but comes with an additional computation cost of calculating the superpixels compared to rectangular patch classification approaches. However, compared to traditional rectangular patch based algorithm, the proposed algorithm yields better performance, as boundary information is respected during partitioning the image into superpixels. The empirical evaluation of the multi-scale superpixel classification algorithm shows that it yields very high classification performance in terms of area under ROC curve. Although the evaluation of the performance of our algorithm has been on a breast tissue dataset, the technique described here can, in essence, be applied to other tissue types as well. Moreover, the multi-resolution superpixel classification approach can potentially be utilized to discriminate between cancerous and normal regions. This will be the subject of future work.

Acknowledgment

The authors wish to acknowledge the financial support by the European Union FP7 funded VPH-PRISM project under grant agreement n°601040.

Automated detection of DCIS in whole-slide images of breast tissue

5

Babak Ehteshami Bejnordi, Maschenka Balkenhol, Geert Litjens, Roland Holland, Peter Bult, Nico Karssemeijer, and Jeroen AWM van der Laak

Original title: Automated detection of DCIS in whole-slide H&E stained breast histopathology images

Published in: IEEE Transactions on Medical Imaging, 35(9), 2141-2150, 2016.

Abstract

This paper presents and evaluates a fully automatic method for detection of ductal carcinoma in situ (DCIS) in digitized hematoxylin and eosin (H&E) stained histopathological slides of breast tissue. The proposed method applies multi-scale superpixel classification to detect epithelial regions in whole-slide images (WSIs). Subsequently, spatial clustering is utilized to delineate regions representing meaningful structures within the tissue such as ducts and lobules. A region-based classifier employing a large set of features including statistical and structural texture features and architectural features is then trained to discriminate between DCIS and benign/normal structures. The system is evaluated on two datasets containing a total of 205 WSIs of breast tissue. Evaluation was conducted both on the slide and the lesion level using FROC analysis. The results show that to detect at least one true positive in every DCIS containing slide, the system finds 2.6 false positives per WSI. The results of the per-lesion evaluation show that it is possible to detect 80% and 83% of the DCIS lesions in an abnormal slide, at an average of 2.0 and 3.0 false positives per WSI, respectively. Collectively, the result of the experiments demonstrate the efficacy and accuracy of the proposed method as well as its potential for application in routine pathological diagnostics. To the best of our knowledge, this is the first DCIS detection algorithm working fully automatically on WSIs.

5.1 Introduction

Breast cancer is the second leading cause of cancer death among women[83]. Approximately 80% of breast cancers arise from epithelial cells lining the ducts (ductal carcinoma). Pathological diagnosis for intraductal proliferative lesions comprise a spectrum with increasing malignant potential, ranging from usual ductal hyperplasia (UDH), atypical ductal hyperplasia (ADH), ductal carcinoma in situ (DCIS), to invasive ductal carcinoma (IDC)[84]. In this spectrum, DCIS (with cancer cells still being contained within the glandular tissue) and IDC (cancer cells invading the surrounding tissue) are considered malignant, prompting for immediate treatment[83].

DCIS encompasses a heterogeneous group of lesions with highly variable morphology, biomarker expression, genomic profile, and natural progression[85]. Whereas the extremes of the spectrum are easily discernible, the difference between UDH, ADH, and low-grade DCIS is subtle and the classification of such lesions suffers from significant inter-observer variability even among expert pathologists. Introduction of computer aided diagnosis (CAD) systems for breast pathology will be successful if these difficult cases can be handled, with sufficient accuracy. CAD can assist the pathologist in two ways: (1) by detecting all the clinically relevant regions of interest (ROIs) per slide, allowing the pathologist to only focus on the interpretation of these regions, or (2) by providing an accurate assessment of suspicious regions and reducing the variability in pathologists' interpretations. Several recent studies have focused on automated discrimination of DCIS from benign intraductal breast lesions[86–88]. Two approaches, based on identification and segmentation of nuclei and the quantification of nuclear features by Dong et al.[86] and Dundar et al.[87] could discriminate DCIS from UDH with area under the receiver operating characteristic curve (AUC) of 0.86 and 0.93, respectively. Srinivas et al.[88] proposed a simultaneous sparsity model to automatically evaluate intraductal breast lesions for cancer diagnosis.

One of the major drawbacks of many published studies on CAD in pathology is the fact that only manually selected ROIs (mostly selected by expert pathologists) were used. A fully automated algorithm that can be used in large-scale histopathological image analysis should automatically identify ROIs in the whole-slide-image (WSI) and discriminate DCIS from different types of benign lesions. This task is particularly challenging for two main reasons: (1) WSIs are large and may contain hundreds of structures which need to be analyzed. Therefore, obtaining a small false positive rate while still retaining a high sensitivity can be hard, and (2) A CAD system that operates on the WSI level should be able to handle a larger set of heterogeneous benign structures (e.g. adenosis, UDH, cysts, etc.) and artifacts (due to

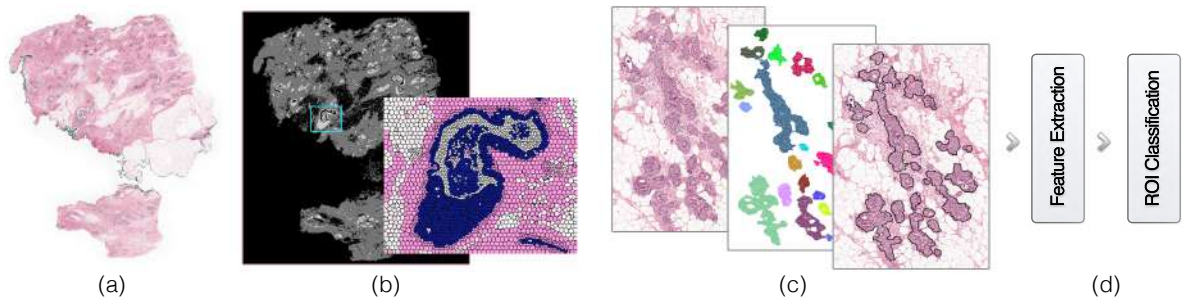

Figure 5.1: Overview of the proposed DCIS detection system. (a) Original WSI of a breast tissue slide. (b) Application of multi-scale superpixel classification to classify the image into epithelium, stroma, and background. (c) Graph-based clustering of the epithelium labeled superpixels for delineation of ROIs. (d) Feature extraction and classification of each of the candidate ROIs.

staining/cutting) to detect DCIS.

In this paper, we present a fully automated CAD system that can discriminate DCIS from normal/benign conditions in WSI. Our proposed system initially detects epithelial regions in the WSI.A common approach to localize important structures in WSIs is to divide the image into rectangular patches and classify them (possibly at multiple resolutions)[74,76,89]. However, these rectangular patches may contain mixtures of class types which will lower the accuracy of the classification. To tackle this problem, our system uses a multi-scale superpixel classification approach[90] to detect epithelial regions in the WSI. Superpixels are classified at multiple resolutions to efficiently detect regions containing epithelium. The superpixels labeled as epithelium are then grouped into histopathologically meaningful regions by application of a graph clustering algorithm. A set of texture and spatial distribution features is then extracted from each candidate region, after which a classifier classifies the region as either DCIS or benign/normal.

Empirical evaluation of the performance of the proposed system is presented in two experiments using two separate datasets. The first dataset comprises 150 WSIs of breast tissue sampled from 150 patients (75 benign/normal and 75 containing DCIS). The second comprises 55 WSIs of breast tissue sampled from 43 patients which are representative of the daily clinical routine samples examined by a pathologist during a specific period of time. The first experiment evaluates the efficacy of the proposed system in detecting and localizing DCIS regions in WSIs using the first dataset. The Dice coefficient is used to evaluate the accuracy of the DCIS segmentation. The second evaluates the performance of the system in classifying a WSI as DCIS at the slide level using the second dataset. This is an important aspect in evaluating the merit of the proposed CAD system because it highlights its potential for

application in routine pathological diagnostics.

5.2 Methods

5.2.1 Detailed description of the proposed DCIS detection system

The proposed DCIS detection system takes as input an H&E stained WSI and yields as output the segmentation of the potential DCIS lesions together with a likelihood estimation for each lesion to be DCIS. Figure 5.1 presents an overview of the proposed DCIS detection system. The proposed algorithm has 3 basic steps:

(a) Multi-scale superpixel classification to find epithelial areas in the WSI.

(b) Graph-based clustering of the superpixels labeled as epithelium and delineation of ROIs.

(c) Classification of the segmented regions as benign/normal or DCIS.

Detailed description of the proposed algorithm's steps are discussed below.

5.2.2 Multi-scale superpixel classification to find epithelial areas

Detection of epithelial regions in the WSI is based on the multi-scale superpixel classification algorithm[90]. This algorithm enables subdivision of the WSI into regions which adapt to the underlying image data, such that every superpixel is mostly homogeneous. Accurate classification of the tissue components within the WSI is thereby facilitated. The algorithm initially partitions the image at 1.25X magnification (with pixel size of $3.88\mu m \times 3.88\mu m$) into a set of non-overlapping superpixels using the simple linear iterative clustering (SLIC) algorithm[79]. The generated superpixels each contain approximately 5000 pixels. Image regions containing mainly epithelium or stroma are identified by excluding all the superpixels whose content is more than 90% background. A pixel within a superpixel is classified as background if its overall optical density (i.e. $-\log(\frac{I}{I_0})$ where I_0 denotes the intensity of the light source) is lower than 0.2 and the density of its r, g, and b channels is lower than 0.25.

Figure 5.2 presents the next steps in the multi-scale superpixel classification algorithm. New superpixels are constructed at 5X magnification (pixel size of $0.97\mu m \times 0.97\mu m$) within the areas classified as epithelium or stroma in the previous step (see Figure 5.2(b)). These are then again classified into three distinct components: stroma, epithelium, and background (non-tissue containing regions as well as regions containing fat cells and fluid). The size of each superpixel at this magnification was

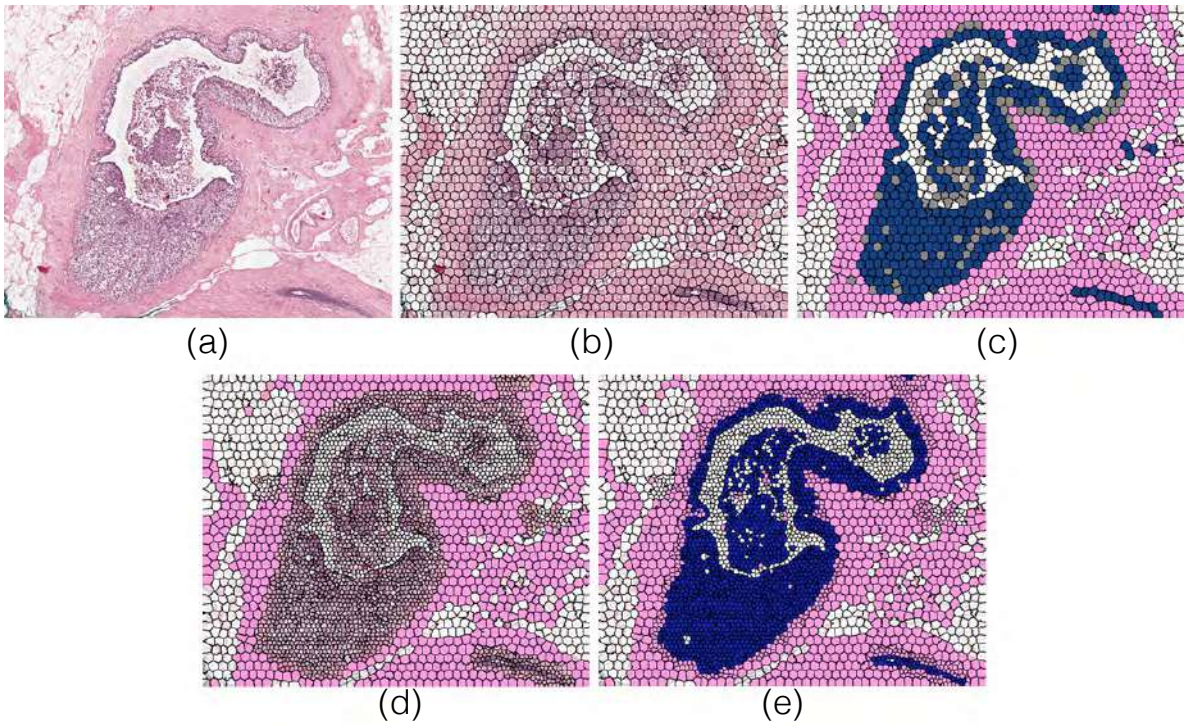

Figure 5.2: Illustration of multi-scale superpixel classification. (a) Original Image. (b) Generation of superpixels in the intermediate magnification (5X). (c) Classification of the superpixels into background (white), stroma (pink), or epithelium (blue). Note that gray-colored superpixels are the ones for which the probability score of the classifier for all of the classes was below 0.7. (d) Generating superpixels on the areas requiring more detailed information in higher magnification (20X). Note that the smaller superpixels are built on the highest magnification image while the larger ones are the same superpixels computed in the intermediate magnification. (e) Final classification result.

set to be approximately 2000 pixels. Identification of background superpixels is performed similarly to the previous step. For the classification of the remaining superpixels into stroma or epithelium, a set of 54 features were extracted for each superpixel s, including 8 pixel value statistics (minimum, maximum, sum, mean, standard deviation, lower quartile, median, and upper quartile) and 10 uniform local binary pattern features for radius 1 derived from each of the channels of the hue-saturation-density (HSD) color model[46]. In addition, the mean and standard deviation of all of these features for the set of all neighboring superpixels to the superpixel s was included, yielding a total of 162 features. A random forest classifier using 100 decision trees trained on approximately 20,000 manually annotated superpixels (generated on sample patches taken from 30 WSIs in the training set) was employed for classifying the superpixels. Figure 5.2c shows the results of the classification at the intermediate magnification.

Finally, to achieve a more accurate delineation of histopathological structures, a new set of superpixels was built and classified at the highest magnification (20X) within the areas requiring more detailed information (see Figure 5.2(d)). A region required more accurate classification (using the highest magnification image) if it satisfied either of these two conditions: (1) The classifier used to classify the region at the intermediate magnification yielded a low confidence in assigning the output label. The level of uncertainty was assigned according to the output probability for the superpixel classification. Superpixels having a likelihood probability lower than 0.7 for all of the classes were considered uncertain. (2) The region was labeled as epithelium by the classifier in the intermediate magnification. The first condition ensures that a more accurate classification is achieved by using more detailed information present in the higher magnification. The second is to obtain more detailed contouring of the areas that were labeled as epithelium.

The newly generated superpixels in the corresponding areas satisfying the two conditions mentioned above had an approximate size of 1000 pixels. The set of 54 features described previously were extracted for each superpixel in this magnification. Moreover, to incorporate more contextual information for the superpixel s, the set of 162 features previously computed for the parent superpixel s' in the intermediate magnification was appended to the feature list, where s' is the superpixel which has the largest overlap with the corresponding area occupied by the superpixel s. A second stage random forest classifier with 100 decision trees was subsequently utilized to classify these superpixels more accurately. Figure 5.2(e) shows the final classification result by the multi-scale superpixel classification approach.

5.2.3 Graph-based clustering of superpixels for delineation of ROIs

The output of the multi-scale superpixel classification algorithm is a set of superpixels with three possible labels (stroma, background, and epithelium). To create regions representing anatomically meaningful structures within the tissue such as ducts or lobules, the superpixels have to be clustered. The aim of this step is not only to merge the superpixels neighboring each other but also splitting distinctive structures lying in the vicinity of each other. To perform the clustering, we propose an algorithm based on local graph structure that models the spatial distribution of the labeled superpixels in the image. Our proposed spatial clustering algorithm prunes the edges of a region adjacency graph built on the centroids of the epithelium-labeled superpixels to cluster them into meaningful tissue regions in the WSI, while still maintaining the overall connectivity of each cluster. The entire algorithm for delineating ROIs can be summarized in three major steps:

1. Step 1: Using a relative neighborhood graph to identify coarse clusters of neighboring superpixels.
2. Step 2: Applying spatial clustering to find spatially homogenous sub-clusters within clusters from the first step.
3. Step 3: Finding the concave-hull of each sub-cluster as the outer boundary of the identified ROI.

Detailed description of each of the steps is discussed below.

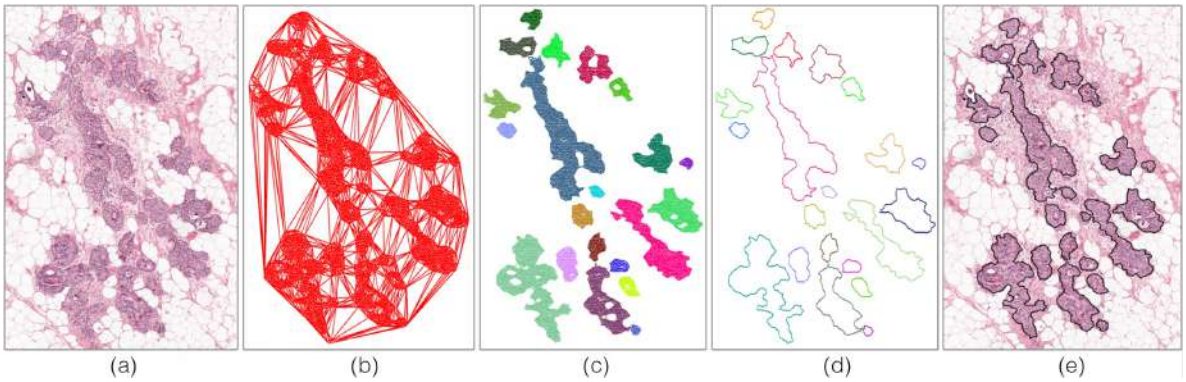

Figure 5.3: Graph-based clustering of superpixels for delineation of ROIs. (a) Original Image. (b) Delaunay triangulation built on the set of epithelium labeled superpixels. (c) Application of the graph-clustering algorithm to cluster the graph into several meaningful sub-graphs. (d) Calculating the concave hull for each of the sub-graphs. (e) Final contouring of each ROI.

Step 1: Identifying coarse clusters

The algorithm for identifying several isolated groups of coarse clusters can be described as follows:

(a) Apply the multi-scale superpixel classification algorithm to the input WSI to obtain labeled superpixels as described in section 5.2.2.

(b) Construct the relative neighborhood graph $RNG(V)$[91] of the pointset V containing the centroids of the n superpixels with epithelium label in Euclidean space. In the relative neighborhood graph two points v_i and v_j are neighbors if $d(v_i, v_j) \leq max\,[d(v_i, v_k), d(v_j, v_k)]$, $\forall k = 1, ..., n$ and $k \neq i, j$.

(c) Apply a threshold ($T = 200$, equivalent to the diameter of two superpixels) on the maximum edge length of the graph to partition $RNG(V)$ into several local sub-graphs (G^k) and label the entire group of subgraphs using the depth-first search (DFS) algorithm[92].

The threshold on maximum edge length was determined based on the assumption that two superpixels lying further than the diameter of two superpixels away should not be considered neighbors.

Step 2: Spatial clustering of coarse clusters into anatomically meaningful sub-clusters

The identified coarse clusters in the previous step may contain multiple anatomically meaningful structures which are lying in the vicinity of each other. In this step we cluster each of the identified $G^k = (N, E)$ into several anatomically meaningful tissue regions such as ducts or lobules, where N and E are the set of vertexes and edges of G^k respectively. Figure 5.3 shows the processes involved in the proposed spatial clustering algorithm. Our spatial clustering algorithm utilizes Delaunay triangulation (DT), which is a suitable tool for spatial clustering as it implicitly encapsulates vast amount of proximity information (see Figure 5.3(b)). The proposed clustering algorithm eliminates the extra simplexes of the triangulation according to a local heterogeneity measure. A DT built on a cloud of points may have both inter- or intra-cluster simplexes. Inter-cluster simplexes are the ones connecting two or multiple anatomically meaningful regions (e.g. ducts or lobules) to each other. Intra-cluster simplexes, however, are the ones connecting multiple vertexes inside a single sub-cluster. Our objective is to extract measures from each simplex in a constructed DT to discriminate between inter- and intra-cluster simplexes and consequently identify separate clusters of points each belonging to separate regions. For this purpose, we define three measures to describe spatial heterogeneity of the simplexes.

The perimeter of the simplex is taken as the first measure describing local topography of DT. According to the density-based definition of clusters, intra-cluster edges are much shorter compared to inter-cluster edges[93,94]. Consequently, it can be inferred that inter-class simplexes have higher perimeter values than the intra-cluster ones.

The second measure quantifies the elongation of the simplex. Inter-cluster simplexes of a DT tend to have more elongated shapes. To measure the elongation $EL(s)$ of the triangle s, we compute the ratio between the major and minor axes of s's Steiner circumellipse[95]. Steiner's Circumellipse is a unique ellipse whose center coincides with the centroid of the triangle and passes through the vertexes of the triangle. For equilateral triangles, the measure $EL(s) = 1$, and for all other conditions $EL(s) > 1$.

Our final measure quantifies the local shape heterogeneity around the simplex. In this way we can evaluate the tendency of the current simplex to be in the same cluster as its neighboring ones. For this purpose, we compute the standard deviation of the

Figure 5.4: Illustration of the selected steps required for computation of some of the architectural features. (a) Original image of a DCIS region. (b) The result of pixel classification using the algorithm proposed in[96]. (c) Euclidean distance transform of the inverse DCIS mask. The mask is computed using the output of the ROI delineation algorithm in step 5.2.3. (d) Portions of *(c)* cut-out by the mask of the hematoxylin stained pixels. These cut-out distances are subsequently used for computation of margination features. (e) Division of the DCIS mask into 10 zones. The ratio between the area of the hematoxylin stained pixels in each zone to the area of that zone are used as measures to characterize the distribution of the nuclei within the candidate ROI.

elongation measures over the set $s \cup N(s)$, where $N(s)$ denotes the set of neighboring simplexes of simplex s. Finally, the entire simplex analysis is captured in a criterion function $F(s)$, which is defined as: $F(s) = Perimeter(s) \times EL(s) \times Std(1 + EL(s \cup N(s))$. This function takes into account spatial heterogeneity of the simplexes and primarily penalizes large simplex perimeters. The two elongation terms are used as weighting factors that further penalize simplexes that have large elongation and/or neighboring simplexes with heterogeneous elongations. For each simplex s in DT, if $F(s)$ is bigger than a predetermined threshold the simplex is removed from the graph. We found the threshold value of 250 suitable. After eliminating inter-class simplexes and noises, only positive nodes and edges of the graph remain. Through depth-first search we then infer the number of isolated clusters and correspondingly the list of points in each cluster.

As a result of pruning the inter-cluster simplexes we may lose the points lying on the hull of each cluster. To reassign these points to the appropriate cluster we start an iterative graph growing process. Let $S^j \subset G^k$ be a clustered graph within the local sub-graph $G^k = (N, E)$, and let $V(i), \{i \in N\}$ denote the set of points neighboring the vertices at the hull of S^j. At each iteration, a point i in $V(i)$ is assigned to S^j under two conditions; (1) if the Euclidean distance between the node i and it's neighboring node in S^j is less than the maximum edge length in S^j. (2) if i is not neighboring the hull of another clustered graph $S^{j'} \subset G^k$. The first rule reduces the possibility of assigning a noisy node to a cluster, and the second will prevent the merging of two isolated clusters. The assignment of new nodes to the graph is repeated for 3

iterations. Because of the two constraints applied, mostly there are no new nodes added to the sub-graphs after 2 or 3 iterations.

Step 3: Finding the outer contour of each sub-cluster

The final step is to extract the boundary of the clustered graphs. Let $G^k = (N, E)$ denote a sub-graph created in step 1 of our algorithm and let $S^j = (N(j), E(j))$ denote a clustered graph obtained in step 2 satisfying $S^j \subset G^k$. To find the actual boundary of the S^j graph which corresponds to the concave hull created by the edges $E(j)$ on the point set $N(j)$ we first compute DT of the point set $N(j)$. The boundary of the exterior face of the DT is the convex hull of the point set. Let Γ_{DT} denote the set of edges of the convex hull. By traversing along the edges of Γ_{DT} and removing and replacing the edges not present in $E(j)$ with the other two edges of the simplexes in DT to which the removed edges belong to, we can find the concave hull of the S^j graph. Traversing along the edges in Γ_{DT} is continued until the condition $\Gamma_{DT} \subset E(j)$ is met. The edges remained in Γ_{DT} correspond to the outer boundary of the S^j graph. At the end of this step, a more accurate delineation of the ROI is obtained by taking the union of the binary masks of the superpixels lying on the concave hull and within the binary mask of the concave hull itself.

5.2.4 Region-based feature extraction and classification

Cellular and architectural features are the major characteristics considered by a pathologist for diagnosing DCIS. Therefore, the features used in this study to distinguish DCIS from different benign/normal regions are a combination of statistical and structural texture features, and features describing the spatial distribution of the components inside the ROI. A classifier is employed to classify each of the segmented ROIs using the extracted features.

Texture features

To extract texture features, each candidate region identified through our spatial-clustering method is first divided into several superpixels having an approximately equal size of 5000 pixels using the SLIC algorithm (the analyzed image has pixel size of $0.486\mu m \times 0.486\mu m$). For each of the superpixels 5 different types of texture features are extracted from the gray-scale intensities of the image. These features are statistics of the gray level histogram (mean, standard deviation, median, first and third quartiles, interquartile range), 14 statistics calculated from the co-occurrence matrix[97], uniform local binary patterns for radii 1 and 2[98], and gray level histogram

statistics extracted from responses to filter banks in particular Laplacian of Gaussian (LoG) at 5 scales, and Gabor filters at 4 scales and 8 orientations are extracted. These texture features have shown strong discriminatory power in characterizing histopathology images[89,99,100]. In total 256 features were extracted for each superpixel. The mean, standard deviation, 5th and 95th percentile of each feature over all superpixels in an ROI yielded a total of 1024 region-based features. Computation of the texture features at the superpixel level rather than pixel level was done to reduce the computation cost of the statistics which are finally derived from them. Moreover, using super-pixels it is possible to extract regions that contain homogeneous tissue structures, therefore the extracted features from these regions tend to be more meaningful and discriminative.

Architectural features

An initial step before computation of the architectural features is classifying the region into different tissue components. For this purpose, we use our recently proposed algorithm[96] for robust stain classification which makes use of spatial information. This algorithm operates at the WSI and automatically extracts training samples for each stain class (the class absorbing mostly hematoxylin and the class absorbing mostly eosin) from the image, obviating the need for manually labeled training data. This algorithm is an intermediate step in the published stain standardization algorithm[96]. Figure 5.4(b) shows an example classification result for a detected ROI. The classified image is median filtered (kernel size 5×5) for removing noisy labels from the result.

Following extraction of the masks for different stain classes, we compute the area of the hematoxylin stained, eosin stained and background pixels. These three area measures implicitly include information about the size of the ROI. For this reason, three additional features were included that were normalized to the total area of the ROI.

A subset of features were designed to measure the margination of the nuclei. The margination features characterize the distances of the nuclei to the ROI boundary. The steps required for computation of the margination features are illustrated in Figure 5.4. First, the Euclidean distance transform of the inverse mask of the ROI is computed as shown in Figure 5.4(c). Next, portions of this image are cut out by the mask of the hematoxylin stained pixels (see Figure 5.4(d)). The distribution of the cut-out distances were quantified at five percentiles (10th to 90th in ten percentile steps). Five additional features were included by normalizing the percentile values to the maximum value of the distance transform of the ROI.

A subset of architectural features have been computed to describe the distribution

of the nuclei within the ROI. To compute these features, the area inside the ROI is first divided into ten different zones $Z_k = \left\{ i \in ROI \mid \frac{D_{max}*(k-1)}{10} < D(i) < \frac{D_{max}*k}{10} \right\}$ where i is an arbitrary pixel inside the ROI, $k \in \{1, 2, 3, ..., 10\}$, $D(i)$ denotes the distance of the pixel i to the boundary of the ROI and D_{max} the maximum distance from the ROI boundary. Figure 5.4(e) shows example of the division of the DCIS mask into 10 zones using this approach. The ratio between the area of the hematoxylin stained pixels in region Z_k to the area of Z_k are defined to characterize the distribution of the nuclei within the ROI.

The final subset of architectural features include three measures to quantify the clustering of background and eosin stained pixels. These features are the maximum of the distance transform of the inverse eosin mask, inverse background mask, and the inverse of the union of the two masks.

Classification of anatomically meaningful regions

The extracted texture and architectural features yielded a total of 1054 features. The performance of three classifiers were evaluated: logistic regression (LR) with L1 regularization ($\lambda = 1$), support vector machine (SVM) with a radial basis function (RBF) kernel (gamma = 10^{-5} and cost = 10^4), and gradient boosted classifier with decision trees (GBC)[101] (with 1000 estimators and learning rate of 0.1). The three classifiers were trained and evaluated on separate training and test sets. The parameters of the classifiers were optimized with cross-validation on the training set. All the parameters of the multi-scale superpixel classification algorithm and the graph-clustering algorithm were defined using a subset of images in the training set.

5.3 Empirical evaluation

5.3.1 WSIs of breast tissue and ground truth

Two image datasets were used in this study for empirical evaluation of the proposed DCIS detection system. The first dataset originates from 150 digitized H&E stained breast tissue slides sampled from 150 patients. Each slide was reviewed independently by two expert breast pathologists (RH and PB) and assigned a pathological diagnosis. 75 of the WSIs contained DCIS (grade I (9), grade II (35), grade III (31)) and 56 contained different types of benign lesions (usual ductal hyperplasia (11), adenosis (8), fibrosis (7), duct ectasia (5), fibrocystic (5), hamartoma (4), pseudoangiomatous stromal hyperplasia (3), sclerosing lobular hyperplasia (2), fibroadenoma (1), and mixed benign lesions (14)) and 19 normals. This dataset was taken from the archives of the department of Pathology. To be able to train and test our algorithm

on different benign lesions that may occur in routine diagnostics, we enriched the benign class with cases containing all types of benign lesions as listed in the national Dutch breast cancer guidelines. Relative occurrence of these lesions in our dataset is comparable to that encountered in routine diagnostics.

The second dataset used in this study is representative of the daily clinical practice of breast pathology examined by a pathologist during a specific period of time. We took all cases from routine diagnostics of one breast pathologist involved in this study (PB) during the period June 2015 to September 2015, containing either DCIS or normal/benign conditions. This dataset consisted of 55 digitized H&E stained breast tissue slides sampled from 43 patients. This dataset contained 20 slides with DCIS diagnosis (grade I (5), grade II (7), grade III (8)) and 35 benign/normal slides (normal (12), calcification (9), usual ductal hyperplasia (5), fat necrosis (5), cyst (4)). Because the second dataset is taken consecutively from our routine diagnostics, in comparison to the first set which is enriched for benign abnormalities, the second set contains much fewer of the various benign lesion categories present in the first set.

In this study, we excluded the slides containing atypical ductal hyperplasia (ADH). The major problem with ADH is the difficulty in achieving acceptable levels of concordance or consistency in diagnosis[102]. Due to the use of different criteria for defining the characteristics of ADH in the literature[103–105] and the difficulty in obtaining reliable ground truth, we chose to exclude this category in our study.

All slides were stained in our laboratory and digitized using the 3DHISTECH Pannoramic 250 Flash II digital slide scanner with a 20X objective lens. Each image has square pixels of size $0.243\mu m \times 0.243\mu m$ in the microscope image plane.

A total of 823 regions containing DCIS in abnormal slides from the first dataset were annotated. All the annotations were verified by two pathologists (RH and PB) independently. We included a lesion as ground truth in case both pathologists were in agreement. No annotation was provided for the slides with benign/normal diagnosis, as the training samples from these slides were automatically extracted using our automatic ROI detection algorithm. The ground truth data for the second dataset is only available at the slide level.

5.3.2 Experiments

To evaluate the performance of the proposed DCIS detection system two experiments were performed. The first experiment evaluates the efficacy of the proposed system in detecting and localizing DCIS regions in WSIs using the first dataset. The second evaluates the performance of the system in classifying a WSI as DCIS at the slide level using the second dataset.

Experiment 1

For this experiment, the first dataset was split into two independent subsets for training and testing. The train set comprises of 50 DCIS slides and 50 benign/normal slides (attempting to balance different DCIS grades and benign lesion categories over train and test sets). The test set comprises 25 DCIS slides and 25 benign/normal slides.

The training samples from the benign/normal category were automatically extracted using the ROI detection and delineation step of our proposed system. The training samples for the DCIS lesions, however, were taken directly from the annotated ROIs.

The performance of the proposed system was evaluated in terms of detecting and localizing the lesion in the slide. A ground truth DCIS lesion was deemed to have been detected if its intersection with the segmentation of the DCIS region performed by the proposed algorithm was non-empty. For the evaluation, free-response receiver operating characteristic (FROC) curve[106] was used. The FROC curve is defined as the plot of sensitivity versus the average number of false positives per image. The FROC curve is computed by varying thresholds on DCIS classification confidence. Considering that not all the DCIS lesions present in abnormal slides may have been annotated, the false positives were only counted in benign/normal slides.

In this experiment, we also evaluate the segmentation performance of the proposed system by computing Dice's overlap measure at the slide level.

Experiment 2

The aim of this experiment was to evaluate the performance of the proposed detection system on an independent dataset representing the daily clinical practice of breast pathology examined by a pathologist. Each of the classifiers was trained independently on the entire slides present in the first dataset and evaluated on the second dataset. The parameters of the classifiers were kept the same as the first experiment.

This experiment evaluates the performance of the system in differentiating between the slides containing DCIS and benign/normal slides. To achieve a slide-based score, the highest scored region in a slide is used as the likelihood score that the case contains DCIS.

Slide-based FROC analysis was performed to evaluate the efficacy of the system. The FROC curve in this experiment plots the fraction of slides classified as DCIS divided by the total number of slides with DCIS versus the average number of false positives per WSI.

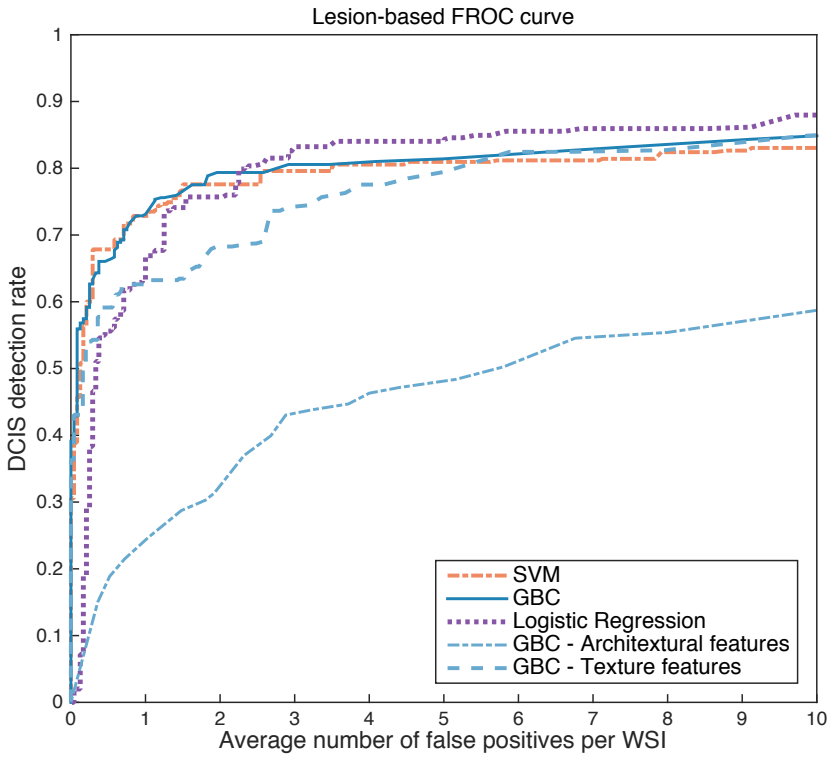

Figure 5.5: Lesion-based FROC curve of the proposed DCIS detection system for experiment 1.

5.3.3 Results

Figure 5.5 presents the FROC curve of experiment 1 for the three classifiers. Note that the false positive rate plotted on the horizontal axis is counted on benign/normal slides only. The FROC curve for the GBC is also presented when only texture features or architectural features were used. Table 5.1 summarizes the DCIS detection (sensitivity) levels at different average numbers of false positives per WSI, for different classifiers. Overall, the three classifiers achieved comparable performance. SVM and GBC yielded higher sensitivities at smaller numbers of false positives while LR performed better at larger number of false positives. Figure 5.6 shows examples of true positives, false positives, as well as false negatives obtained by the CAD system trained by GBC when the performance was fixed at 80% sensitivity.

For the evaluation of the performance of the segmentation algorithm we used Dice's overlap measure. The Dice score at the slide level when considering only the detected DCIS lesions (sensitivity was fixed at 80% in experiment 1) was 0.9243 ± 0.0187 (mean $\pm$ std) over the entire slides in the test set of the first dataset.

In the second experiment GBC yielded the best performance. The FROC curve of Experiment 2 for GBC is shown in Figure 5.7. The 95% confidence interval was generated using patient-stratified bootstrapping with 1000 replications. Table 5.2

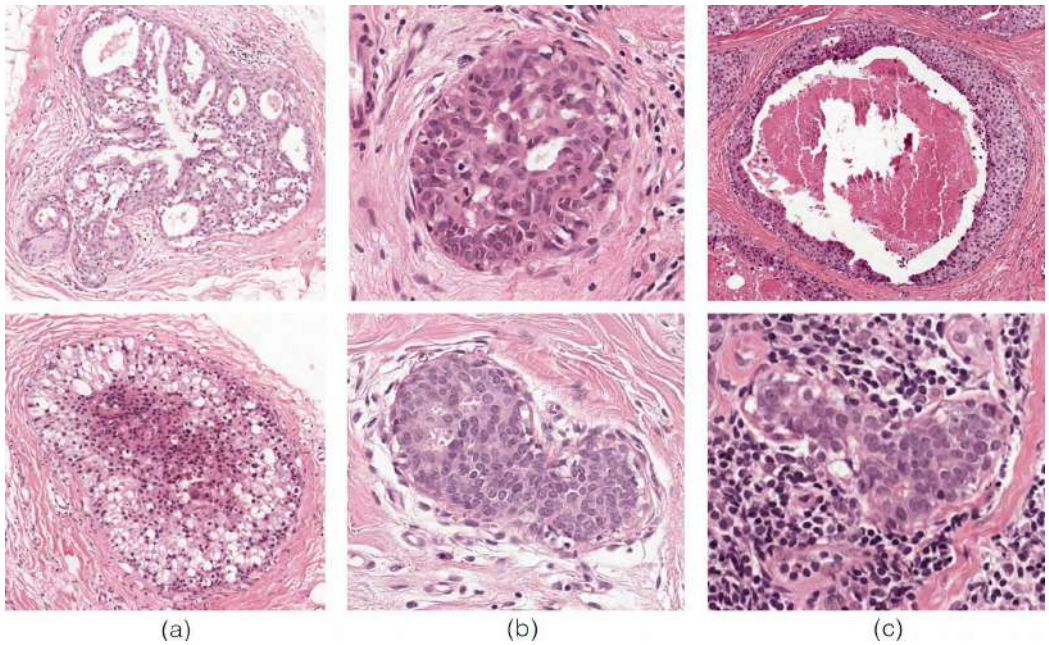

Figure 5.6: Examples of true positives, false positives, as well as false negatives. (a) Shows examples of two correctly detected DCIS lesions. (b) Two false positive examples. (c) Examples of two missed DCIS lesions. The image on top shows a DCIS lesion with large amount of necrosis, and the image in bottom shows an example of a DCIS lesion (lobular cancerization) surrounded by lymphocytes.

Table 5.1: Results of the experiment 1: Sensitivity of DCIS lesion detection is provided at five levels of average numbers of false positives (FPs) per WSI

FPs/WSI	1/2	1	2	3	4
GBC	0.66	0.73	0.80	0.81	0.81
SVM	0.68	0.73	0.78	0.80	0.80
LR	0.50	0.63	0.76	0.83	0.84

summarizes the slide-based DCIS classification sensitivity at different average numbers of false positives per WSI, for different classifiers. Overall, GBC yielded the best performance, achieving a sensitivity of 95% and 100% at average false positive rates of 2 and 2.6, respectively.

5.4 Discussion and conclusion

In this paper, we presented a CAD system for DCIS detection in digitized H&E stained histopathological breast tissue slides. The proposed algorithm is fully automated, does not require any human interaction, and therefore yields objective and reproducible results. Lesion-based and slide-based evaluation of the performance

Table 5.2: Results of the experiment 2: Slide-based sensitivity of DCIS detection is provided at five levels of average numbers of false positives (FPs) per WSI

FPs/WSI	1/2	1	2	3	4
GBC	0.55	0.80	0.95	1.0	1.0
SVM	0.70	0.70	0.85	0.85	0.85
LR	0.55	0.75	0.90	0.95	0.95

of the proposed CAD system was presented. Collectively, the results of the experiments demonstrate the efficacy and accuracy of the proposed CAD system as well as its potential for application in routine pathological diagnostics.

To the best of the authors' knowledge, this is the first fully automated DCIS CAD system that operates at the WSI level and has been evaluated on a dataset collected from routine clinical practice. WSI analysis of histopathological slides remains a challenging medical image analysis problem due to technical complexities in dealing with large WSIs and the requirement to have highly specific algorithms to avoid large numbers of false positives. The focus in developing CAD systems for histopathological images, in particular for the task of recognizing DCIS from different types of benign abnormality, has been mainly limited to analyzing small patch images selected by a pathologist[86–88]. Existing approaches to localize diagnostically relevant regions in WSIs either analyze the WSI at lower magnification or divide the image into rectangular patches and classify them (possibly at multiple resolutions)[74,76,89]. In this study we proposed a multi-scale superpixel classification scheme for finding epithelial areas in WSIs.

Detection and contouring of diagnostically relevant regions was based on a spatial clustering approach operating on the graphs built on the centroids of epithelium labeled superpixels. Several algorithms have been published for the segmentation of glandular structures with application to prostate and colon tissue[107–110]. All of these methods assume an architectural regularity in glandular structure and have detection of lumen as an essential step for segmenting the gland. The intraductal proliferation in the breast, however, usually obliterates and distends the ductal lumen[111] which limits the effectiveness of these methods in detecting DCIS lesions of the breast. The approach proposed by Sirinukunwattana et al.[112] does not have this limitation as it does not necessitate any strict assumption regarding the arrangement of granular components. However, inferring the number of glands in a clustered population of glands is based on the number of isolated connected components resulted from thresholding the glandular probability map. This means partially connected glands may fall in the same connected component and there is no mechanism

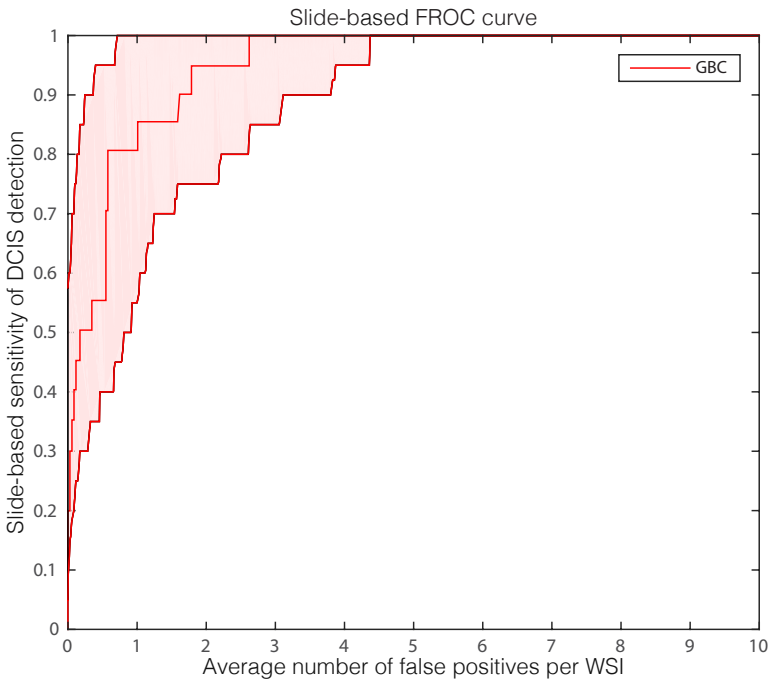

Figure 5.7: Slide-based FROC curve and the 95% confidence intervals of the proposed DCIS detection system for experiment 2.

in the utilized random polygon model (RPM) to further split these glands. Moreover, although the proposed algorithm yields good results in segmenting glands in colon tissue, due to the stochastic modeling nature of RPM, the proposed model has high computational complexity and may not be suitable for application to WSIs. Our proposed spatial clustering algorithm, in contrast, is robust, efficient and well suited for accurate detection and delineation of breast glandular structures in WSIs. Evaluation of the segmentation performance in experiment 1 demonstrate that our spatial clustering algorithm yields a Dice score of 0.9243 ± 0.0187 for segmenting DCIS regions.

Following the segmentation of the diagnostically relevant regions in the WSI, a set of texture-based and architectural features were extracted from the epithelial structure. Figure 5.5 presented the contribution of the proposed architectural features to the performance of the detection system. Our proposed features are efficient to compute and obviate the need to perform nuclear segmentation for describing the distribution of the structures inside the potential DCIS region. Our evaluations on the first dataset demonstrate the efficacy of the proposed method in detecting and localizing DCIS. Using the proposed system, it is possible to detect 80% of the DCIS lesions in an abnormal slide at an average number of 2 false-positive per WSI. Practically, we expect the time gain in automatically detecting 80% of the DCIS lesions in WSIs outweighs the time lost for looking at the false positives.

In this study, we also presented an evaluation on a dataset collected from routine clinical practice during a four month period. This dataset contains large categories of benign lesions that the pathologist encounters in routine diagnostics. Unlike the previous studies which mainly focus on discrimination of DCIS from UDH, the proposed system was designed to handle a large set of heterogeneous benign categories. Evaluation of the slide-based DCIS detection on this dataset shows that at an average number of 2.6 false-positive per WSI, 100% of the slides that contained DCIS could be detected. This suggests the potential of the proposed method for application in routine pathological diagnostics. Moreover, further reduction of the average number of false positives per WSI can significantly reduce the workload of the pathologist as it would mean that a large number of normal/benign slides can be put aside without the risk of missing slides containing DCIS.

The proposed system has several components, of which only a small number may impact the performance. The multi-resolution superpixel classification algorithm utilizes two classifiers trained on manually labeled superpixels. By using our recently proposed algorithm for standardization of WSIs[96], we can obviate the need for re-training these classifiers when applied to new datasets. However, the images in this study were not standardized as the slides were stained using the same protocol and scanned using the same scanner. We have found that the specific choice of parameters for many components of our system such as the threshold applied to determine the uncertainty of the classification of superpixels, the number of iteration for graph growing operation, and the threshold applied for coarse clustering the structures in the WSI, are relatively unimportant, and serve mainly to reduce computational cost. Moreover, the designed architectural features are based on an initial classification of the image into different stain classes. The utilized method is an intermediate step in our whole slide standardization algorithm and as shown in our paper[96] it is very robust against variations in histopathological images.

The computation time for different steps of the proposed system to analyze a WSI is as follows. The multi-scale superpixel classification algorithm for finding epithelial regions on the WSI takes between 20 to 45 minutes depending on the amount of tissue (in particular epithelial tissue) on the slide. The graph-clustering algorithm is very efficient and takes on average less than two minutes to generate segmented ROIs. The feature extraction and classification stage together take on average 10 minutes. The implementation is done in C++ and the experimental platform was a laptop with an Intel Core i7 CPU (2.4 GHZ) and 16 GB of Ram.

Several limitations of the proposed method must be acknowledged. First, the multi-scale superpixel classification algorithm puts lymphocytic infiltrates within the same category as the epithelial class. This may occasionally cause the system

to classify the lesions surrounded by lymphocytes as normal. The major reason for this is that the graph clustering algorithm for delineation of the candidate may result in a region including both lymphocytes and DCIS nuclei, hence polluting the statistics of the DCIS region. Lymphocytes are frequently abundant in benign slides, and less-existent in the annotation of the DCIS regions. A region containing a large number of lymphocytes may therefore be characterized as normal by the system (see Figure 5.6(c) for a false negative). Another limitation of the proposed system is in dealing with lesions having large areas of necrosis and very little epithelium. Figure 5.6(c) shows an example of a DCIS lesion with such characteristics. Possible reasons explaining the difficulty of the proposed system in dealing with these lesions are the lack of training data for such lesions and the significant deviation of the characteristics presented by these lesions compared to the majority of the DCIS lesions.

The proposed system was primarily designed to aid the pathologist in detecting and localizing the lesions in the WSI and giving a second opinion on the malignancy likelihood of the findings. The proposed system, however, has the potential to be applied to related problems, such as detecting and classifying glands in prostate tissue WSIs. In addition, it has provided another important implication for future research. The proposed system can serve as an important first step for development of systems that aim at finding prognostic and predictive biomarkers within malignant lesions, requiring an accurate delineation of such regions. This will be the major direction for future research.

Acknowledgment

The authors wish to acknowledge the financial support by the European Union FP7 funded VPH-PRISM project under grant agreement n°601040. The authors also acknowledge financial support from the Stichting IT Projecten (STITPRO) and the Radboud Institute for Health Sciences (RIHS), both in Nijmegen, The Netherlands. We also gratefully acknowledge the support from the histology laboratory of the Department of Pathology, Radboud University Medical Center, Nijmegen, the Netherlands.

Context-aware stacked convolutional neural networks for classification of breast carcinomas

6

Babak Ehteshami Bejnordi, Guido Zuidhof, Maschenka Balkenhol, Meyke Hermsen, Peter Bult, Nico Karssemeijer, Bram van Ginneken, Geert Litjens and Jeroen AWM van der Laak

Original title: Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images

Accepted for publication in: Journal of Medical Imaging, 2017

Student: Dr. Einstein, Aren't these the same questions as last year's [physics] final exam?

Dr. Einstein: Yes; But this year the answers are different.

Albert Einstein

Abstract

Currently, histopathological tissue examination by a pathologist represents the gold standard for breast lesion diagnostics. Automated classification of histopathological whole-slide images (WSI) is challenging owing to the wide range of appearances of benign lesions and the visual similarity of ductal carcinoma in-situ (DCIS) to invasive lesions at the cellular level. Consequently, analysis of tissue at high resolutions with a large contextual area is necessary.

In this paper, we present context-aware stacked convolutional neural networks (CNN) for classification of breast WSIs into normal/benign, DCIS, and invasive ductal carcinoma (IDC). We first train a CNN using high pixel resolution to capture cellular level information. The feature responses generated by this model are then fed as input to a second CNN, stacked on top of the first. Training of this stacked architecture with large input patches enables learning of fine-grained (cellular) details and global tissue structures. Our system is trained and evaluated on a dataset containing 221 WSIs of H&E stained breast tissue specimens. The system achieves an AUC of 0.962 for the binary classification of non-malignant and malignant slides and obtains a 3-class accuracy of 81.3% for classification of WSIs into normal/benign, DCIS, and IDC, demonstrating its potentials for routine diagnostics.

6.1 Introduction

Breast cancer is the most frequently diagnosed cancer among women worldwide. The most frequent subtype of breast cancer, invasive ductal carcinoma (IDC), accounts for more than 80% of all breast carcinomas. IDC is considered to develop through sequential stages of epithelial proliferation starting from normal epithelium to invasive carcinoma via hyperplasia and ductal carcinoma in situ (DCIS)[113]. DCIS is the pre-invasive stage of breast cancer in which the abnormal cells are confined to the lining of breast ducts. Accurate diagnosis of DCIS and IDC and their discrimination from benign diseases of the breast are pivotal to determine the optimal treatment plan. The diagnosis of these conditions largely depends on a careful examination of hematoxylin and eosin (H&E) stained tissue sections under a microscope by a pathologist.

Microscopic examination of tissue sections is, however, tedious, time-consuming, and may suffer from subjectivity. In addition, due to extensive population-based mammographic screening for early detection of cancer, the amount of data to be assessed by pathologists is increasing. Computerized and computer-aided diagnostic systems can alleviate these shortcomings by assisting pathologists in diagnostic decision-making and improving their efficiency. Computational pathology systems can be used to sieve out obviously benign/normal slides and to facilitate diagnosis by pointing pathologists to regions highly suspicious for malignancy in whole slide images (WSI) as well as providing objective second opinions[25,40].

Numerous efforts have been undertaken to develop systems for automated detection of breast carcinomas in histopathology images[86,87,99,100,114–119]. Most of the existing algorithms for breast cancer detection and classification in histology images involve assessment of the morphology and arrangement of epithelial structures (e.g. nuclei, ducts). Naik et al.[100] developed a method for automated detection and segmentation of nuclear and glandular structures for classification of breast cancer histopathology images. A large set of features describing the morphology of the glandular regions and spatial arrangement of nuclei was extracted for training a support vector machine classifier, yielding an overall accuracy of 80% for classifying different breast cancer grades on a very small dataset containing a total of 21 preselected small regions of interest images. Doyle et al.[99] further investigated the use of hand-crafted texture features for grading breast cancer histopathology images. Dundar et al.[87] and Dong et al.[86] developed automated classification systems based on an initial segmentation of nuclei and extraction of features to describe the morphology of nuclei or their spatial arrangement. While all of the previously mentioned algorithms were designed to classify manually selected regions of interest (mostly

selected by expert pathologists), we[114] proposed an algorithm based on multi-scale analysis of superpixels[120] for automatic detection of ductal carcinoma in situ (DCIS) that operates at the whole slide level and distinguishes DCIS from a large set of benign disease conditions. Recently, Balaz16 et al.[115] proposed a system for detection of regions expressing IDC in WSIs. This system first divides the WSI into a set of homogeneous superpixels, and subsequently, uses a random forest classifier[121] to determine if each region indicates cancer.

Recent advances in machine learning, in particular, deep learning[122,123], have afforded state-of-the-art results in several domains such as speech[124] and image recognition[31,125]. Deep learning is beginning to meet the grand challenge of artificial intelligence by demonstrating human-level performance on tasks that require intelligence when carried out by humans[126]. Obviating the need for domain-specific knowledge to design features, these systems learn hierarchical feature representations directly from data. On the forefront of methodologies for visual recognition tasks are convolutional neural networks (CNN). A CNN is a type of feed-forward neural network defined by a set of convolutional and fully connected layers. The emergence of deep learning, in particular, CNN, has also energized the medical imaging field[36] and enabled development of diagnostic tools displaying remarkable accuracy[127–132]. These motivate the use of CNNs for detection and/or classification of breast cancer in breast histopathology images.

Cruz et al.[116] proposed the first system using a CNN to detect regions of IDC in breast WSIs. In contrast to the modern networks that use very deep architectures to improve recognition accuracy, the utilized network was a 3-layer CNN. Due to computational constraints, the model was only trained to operate on images downsampled by a factor of 16. In a recent publication[117], the authors obtained comparable performance when training and validating their system on a multi-center cohort. Rezae16 et al.[118] trained a multi-stream 5-layer CNN taking as input a combination of RGB images, and magnitude and phase of shearlet coefficients. In all these works, the models were evaluated at the patch-level. In a recent work[119], we demonstrated the discriminating power of features extracted from tumor-associated stromal regions identified by a CNN for classifying breast WSI biopsies into invasive or benign.

Different from the above mentioned approaches, the aim of the present study is to develop a system for WSI classification of breast histopathology images into three categories: normal/benign, DCIS and IDC categories. This problem is particularly difficult because of the wide range of appearances of benign lesions as well as the visual similarity of DCIS lesions to invasive lesions at the cellular level. Figure 6.1 shows some examples of lesions in our dataset. A system capable of discriminating

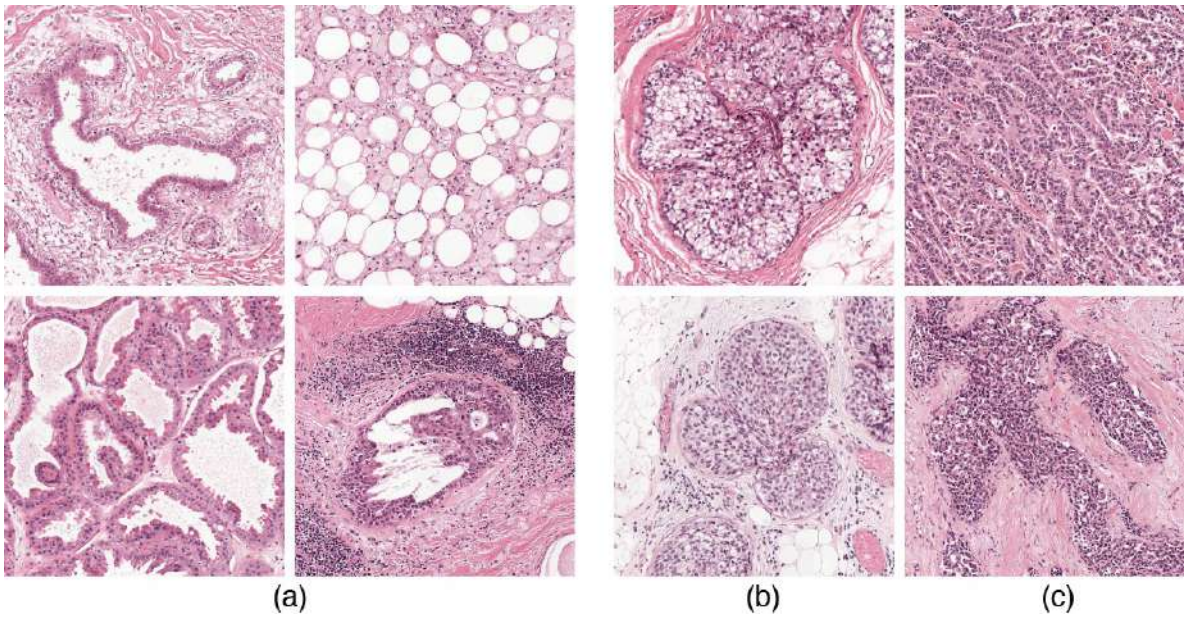

Figure 6.1: Example of breast tissue strutures/lesion. (a) Normal tissue and benign lesions. (b) Ductal carcinoma in situ (DCIS). (c) Invasive ductal carcinoma (IDC).

these three classes, therefore, needs to use high-resolution information for discriminating benign lesions from cancer along with contextual information to discriminate DCIS from IDC. To develop a system that will work in a clinical setting, this study uses WSIs rather than manually extracted regions. Also, the cases in the 'non-malignant' category contained many of the common benign lesions, as they appear in pathology practice.

To this end, we introduce context-aware convolutional neural networks for classification of breast histopathology images. First, we use a deep CNN which uses high pixel resolution information to classify the tissue into different classes. To incorporate more context to the classification framework, we feed a much larger patch to this model at test time. The feature responses generated by this model are then input to a second CNN, stacked on top of the first. This stacked network uses the compact, highly informative representations provided by the first model, which, together with the information from surrounding context, enables it to learn the global interdependence of various structures in different lesion categories. The performance of our system is evaluated on a large breast histopathology cohort comprising 221 WSIs from 122 patients.

6.2 Methods

6.2.1 Overview of the system

The main challenge in the design of our classification framework is that the appearance of many benign diseases of the breast (e.g. usual ductal hyperplasia) mimic that of DCIS, hence requiring accurate texture analysis at the cellular level. Such analysis, however, is not sufficient for discrimination of DCIS from IDC. DCIS and IDC may appear identical on cellular examination but are different in their growth patterns which can only be captured through the inclusion of larger image patches containing more information about the global tissue architecture. Because of computational constraints, however, it is not feasible to train a deep CNN with large patches at high resolution that contain enough context.

Our method for classification of breast histopathology WSIs overcomes these problems through sequential analysis with a stack of CNNs. The key components of our classification framework, including the CNN used for classification of high-resolution patches, the stacked CNN for producing dense prediction maps, and a WSI labeling module are detailed in the following sections.

6.2.2 Deep CNN for classification of small high-resolution patches

Inspired by the recent successes of deep residual networks[133] for image classification, we trained and evaluated the performance of this CNN for classification of small high-resolution patches. We applied an adaptation of the ResNet architecture called wide ResNet as proposed by Zagoruyko et al.[134]. This architecture has two hyperparameters: N and K determining the depth and width of the network. We empirically chose $N = 4$ and $K = 2$ to as a trade-off between model capacity, training speed and memory usage. Hereafter, we denote this network as *WRN-4-2* (see Figure 6.2). This network takes as input patches of size 224×224. Zero padding was used before each convolutional layer to keep the spatial dimension of feature maps constant after convolution.

The goal of this step was to transfer the highly informative feature representations learned by this network produced at its last convolutional layer to a stacked network which is described next.

6.2.3 Context-aware Stacked CNN (CAS-CNN)

In order to increase the context available for dense prediction, we stack a second CNN on top of the last convolutional layer of the previously trained *WRN-4-2* net-

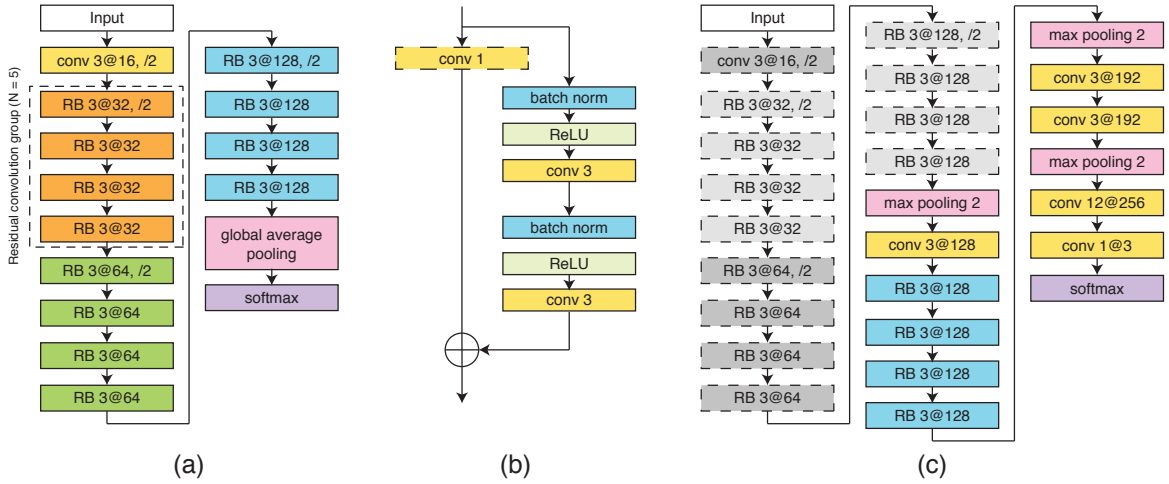

Figure 6.2: Architectures used for patch classification. (a) The *WRN-4-2* architecture used for classification of 224×224 input patches. This architecture consists of an initial convolutional layer that is followed by three residual convolution groups (each of size N=4 residual blocks), followed by global average pooling and a softmax classifier. Downsampling is performed by the first convolutional layers in each group with a stride of 2 and the first convolutional layer of the entire network. Here, *Conv 3@32* is a convolutional layer with a kernel size of 3×3, and 32 filters. (b) The Residual Block (RB) used in this paper. Batch normalization and ReLU precede each convolution. $\oplus$ indicates an element-wise sum. Note that the 1×1 convolution layer is only used in the first convolutional layer of each Residual convolution group. (c) Architecture of the CAS-CNN with input size of 768×768. The weights of the components with dotted outlines are taken from the previously trained *WRN-4-2* network, and are no longer updated during training.

work. The architecture of the stacked network, as shown in Figure 6.2, is a hybrid between the wide ResNet architecture and the VGG architecture[135]. CAS-CNN is fully convolutional and enables fast dense prediction due to re-using of overlapping convolutions during inference. All the parameters of the *WRN-4-2* network were fixed during training. Despite being trained with fixed input patches of size 224×224, because of being a fully convolutional network, *WRN-4-2* can take a larger patch size during training of the stacked network, and consequently produce feature maps with larger spatial dimensions. Moreover, because of fixing the parameters of *WRN-4-2*, the intermediate feature maps of this network do not need to be stored during backpropagation of the gradient. This allowed us to train stacked networks with much larger effective patch sizes. Consequently, we trained 3 networks with patch sizes of 512×512, 768×768, and 1024×1024.

Producing the dense prediction for a given WSI involved sliding the stacked network over the WSI with a stride of 224.

6.2.4 WSI labeling

Given a prediction map produced by the stacked network, we extracted a set of features describing global information about the lesions and their architectural distribution for subsequent classification into normal/benign, DCIS, or IDC. To this end, the probability map was transformed into a three label map, by assigning the class with the highest probability for every pixel. The three label map could contain several connected components for different object classes which were used for extracting features. Next, we describe the set of features extracted for WSI labeling.

Global lesion features

These include the fraction of pixels classified as benign, DCIS, IDC, or cancerous (DCIS and IDC combined) with respect to all non-background pixels, along with the fraction of DCIS, and IDC labeled pixels with respect to all cancerous pixels. We additionally computed a convex-hull area feature for IDC detected lesions. IDC lesions usually appear as a large connected mass. As such, we constructed a convex hull of all IDC detected connected components in the WSI and computed the area ratio between the pixels labeled as IDC and the area of the convex hull. In case multiple tissue sections were present in the WSI, we took the average of these measures over different tissue sections. Note that IDC labeled connected components with an area smaller than $1500\mu m^2$ were discarded as false positives prior to computation of the convex hull feature. At the end, we computed the average area of DCIS connected components as well as IDC connected components as our two final global features.

Architectural features

These features describe the spatial distribution of DCIS and IDC lesions in the WSI. They were extracted from the area-Voronoi diagram[136] and Delaunay triangulation (DT). We built these graphs for DCIS and IDC lesions independently. The seed points for constructing the graphs were the center of the connected components representing DCIS or IDC lesions.

The set of features computed for each area-Voronoi region includes area, eccentricity, the area ratio of the Voronoi region and the total tissue area, and the area ratio of the lesion inside the Voronoi region and the Voronoi region itself. Per WSI, we computed the mean, median and standard deviation of these Voronoi area metrics. Additionally, we added the area of the largest Voronoi region to the feature set.

The features extracted for each of the nodes in the DT include the number of neighbors that are closer than a certain threshold to the node (threshold = 1500μm), and the average distance of these neighbors to the node. We computed the mean, median and standard deviation of these values as features. Additionally, we added the highest average node distance in the DT to the feature set.

Overall, a total of 57 features were extracted, which were used as input to two random forest classifiers[121] with 512 decision trees: one for 3 class classification of WSIs and the other for binary classification of the WSIs into normal/benign versus cancerous (DCIS and IDC). We tuned the parameters of the classifiers by cross-validation on the combined set of train and validation WSIs.

6.3 Experiments

6.3.1 Data

We conducted our study on a large cohort comprising 221 digitized WSIs of H&E stained breast tissue sections from 122 patients, taken from the pathology archive. Ethical approval was waived by the institutional review boards of the Radboud University Medical Center because all images were provided anonymously. All slides were stained in our laboratory and digitized using the 3DHISTECH Pannoramic 250 Flash II digital slide scanner with a 20X objective lens. Each image has square pixels of size $0.243\mu m \times 0.243\mu m$ in the microscope image plane.

Each slide was reviewed independently by a breast pathologist (PB) and assigned a pathological diagnosis. Overall, the dataset contains 100 normal/benign, 69 DCIS, and 55 IDC WSIs. Two human observers (MB and MH) annotated DCIS and IDC lesions using the Automated Slide Analysis Platform (ASAP)[62]. All the annotations were verified by the breast pathologist. Note that the slide labels were assigned

according to the worst abnormality condition in the WSI. Therefore, a slide with the IDC label may contain both IDC and DCIS lesions.

We split this cohort into three separate sets: one for fitting classification models, one for intermediate validation and model selection, and one set for final evaluation of the system (test set). The training, validation, and test sets had 118 (50 normal/benign, 38 DCIS, and 30 IDC), 39 (19 normal/benign, 11 DCIS, and 9 IDC), and 64 (31 normal/benign, 20 DCIS, and 13 IDC) WSIs, respectively. There was no overlap at the slide- and patient-level between the three sets. The benign/normal category included 15 normal and 85 benign WSIs comprising fibroadenoma (14), ductal hyperplasia (11), adenosis (8), fibrosis (8), fibrocystic disease (8), duct ectasia (7), hamartoma (7), pseudo angiomatous stromal hyperplasia (5), sclerosing lobular hyperplasia (5), and mixed abnormalities (12). The WSIs from these 10 benign categories and the normal class were proportionally distributed in the training, validation, and test sets. Note that the relative occurrence of these lesions in our dataset is comparable to that encountered in routine diagnostics.

6.3.2 Training protocols for CNNs

We preprocessed all the data by scaling the pixel intensities between 0 and 1 for every RGB channel of the image patch and subtracting the mean RGB value that was computed on the training set. The training data was augmented with rotation, flipping, and jittering of the hue and saturation channels in the HSV color model.

Patches were generated on-the-fly to construct mini-batches during training and validation, by random selection of samples from points inside the contour of annotations for each class. For each mini-batch, the number of samples per class was determined with uniform probabilities.

Both WRN-4-2 and CAS-CNN were trained using Nesterov accelerated gradient descent. The weights of all trainable layers in the two networks were initialized using He initialization[137]. Initial learning rates of 0.05 and 0.005 were used for WRN-4-2 and CAS-CNN, respectively. The learning rates were multiplied by 0.2 after no better validation accuracy was observed for a predefined number of consecutive epochs which we denote as epoch patience (E_p). The initial value for E_p was set to 8 and increased by 20% (rounded up) after every reduction in learning rate. We used a mini-batch size of 22 for the WRN-4-2 and 18 for the CAS-CNN trained with patches of size 512×512 and 768×768. The network trained on 1024×1024 patches had a greater memory footprint and was trained with mini-batches of size 10.

Training of the WRN-4-2 involved one round of hard negative mining. Unlike the

annotation of DCIS and IDC regions, the initial manual annotation of normal/benign areas was based on an arbitrary selection of visually interesting areas (e.g. areas that visually resembled cancer). These regions are not necessarily difficult for our network. In addition, some of the more difficult to classify benign regions could be underrepresented in our training set. We, therefore, enriched our training dataset by automatically adding all false positive regions in normal/benign training WSIs resulted by our initially trained WRN-4-2 model.

6.3.3 Empirical evaluation and results

We evaluated the performance of our system for classifying the WSIs into normal/benign, DCIS, and IDC categories using the accuracy measure and Cohen's kappa coefficient[138]. We additionally measured the performance of our system for the binary classification of normal/benign versus cancer (DCIS and IDC combined) WSIs.

As an intermediate evaluation, we began with measuring the performance of the WRN-4-2 for the binary and 3-class problems at the patch level (see in table 6.1). These are only results on the validation set, as this network is not used for producing the dense prediction maps individually. As can be seen, the model performs significantly better for the two class problem with an accuracy of 0.924 compared to the three class accuracy of 0.799 for the 3-class problem. This could be explained by the fact that WRN-4-2 only operates on small patches of size 224×224 and does not have enough context for a more accurate discrimination of the three classes.

Table 6.1: Patch-level accuracy for different networks on the validation set

CLASSIFICATION	PATCH SIZE	ARCHITECTURE	ACCURACY
Normal/Benign, Cancer	224×224	WRN-4-2	0.9241
Normal/Benign, DCIS, IDC	224×224	WRN-4-2	0.7995
Normal/Benign, DCIS, IDC	512×512	CAS-CNN	0.8797
Normal/Benign, DCIS, IDC	768×768	CAS-CNN	0.9050
Normal/Benign, DCIS, IDC	1024×1024	CAS-CNN	0.9135

The results for the performance of the CAS-CNN on the validation set for the 3-class problem are shown in Table 6.1. The 3-class accuracy of this network was considerably improved compared to that of the WRN-4-2 at the patch level. We also observe that increasing the training patch-size leads to better performance. Accuracies of 0.872, 0.905, and 0.914 were obtained for the CAS-CNN networks trained on 512×512, 768×768, and 1024×1024 patches, respectively.

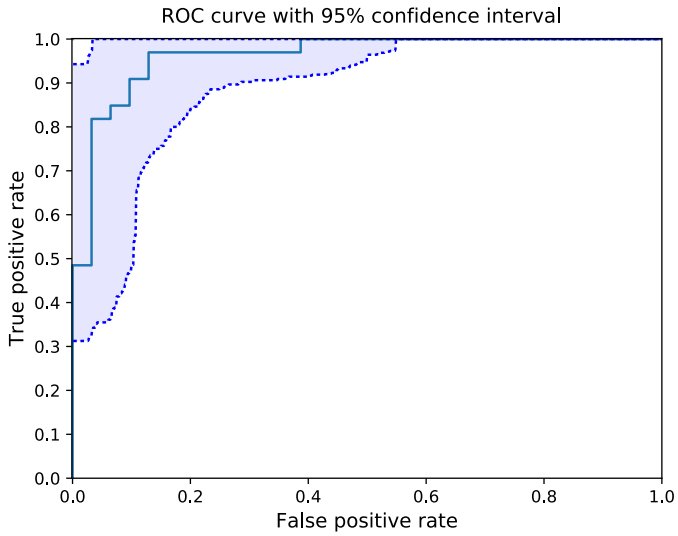

Figure 6.3: ROC curve of the proposed system for binary classification of the WSIs in the test set into normal/benign and cancer (DCIS and IDC).

Due to heavy computational costs of the network operating on 1024×1024 patches, the CAS-CNN network trained on 768×768 was ultimately selected for producing dense prediction maps. The results of the random forest classifier for WSI classification on the test set of our dataset are presented in table 6.2. For the binary classification task, our system achieves an AUC of 0.962. The accuracy and kappa values were 0.891 and 0.781, respectively. The ROC curve of the system for binary classification of WSIs into cancer versus normal/benign is shown in Figure 6.3.

Table 6.2: Results of whole-slide image label prediction on the test set

LABELS	ACC	KAPPA	AUC
Benign, Cancer	0.891	0.781	0.962
Benign, DCIS, IDC	0.813	0.700	-

The system achieves an overall accuracy and kappa value of 0.813 and 0.700 for three class classification of WSIs. The confusion matrix of the test set predictions is presented in table 6.3. Figure 6.4 presents several examples of correctly and incorrectly classified image patches for different lesion classes.

6.4 Discussion and conclusion

In this paper, we presented a context-aware stacked CNN (CAS-CNN) architecture to classify breast WSIs. To the best of our knowledge, this is the first approach investigating the use of deep CNNs for multi-class classification of breast WSIs into

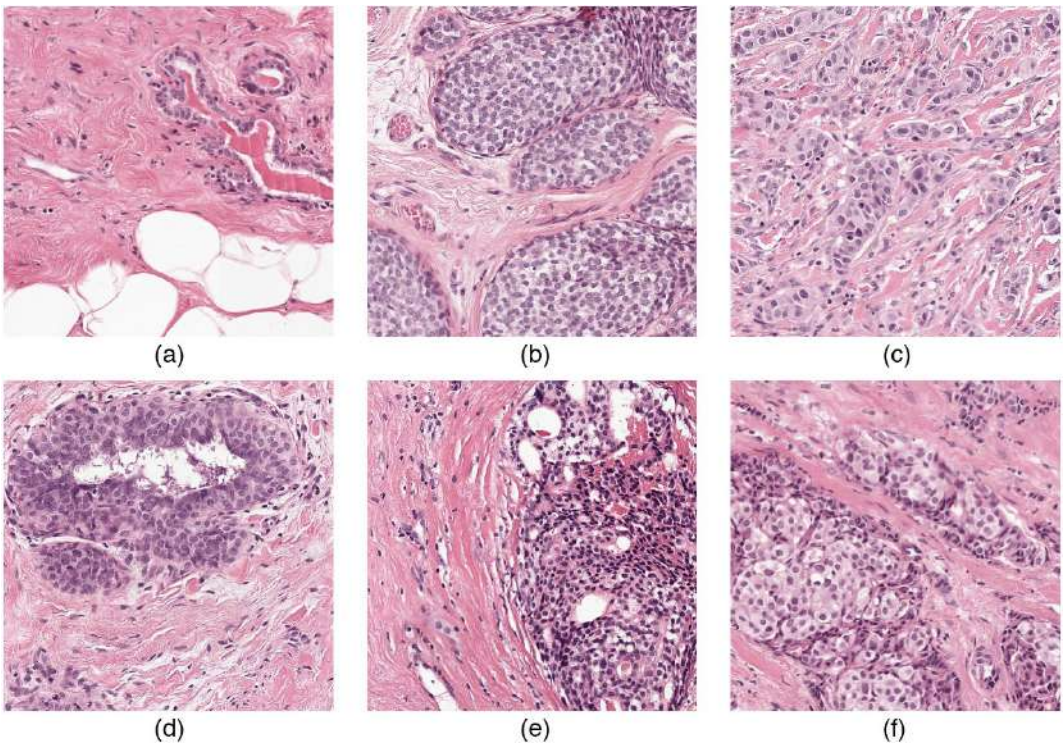

Figure 6.4: Examples of correctly and incorrectly classified patches for different types of lesions. (a-c) correctly classified normal, DCIS, and IDC regions, respectively. (d) a benign lesion (usual ductal hyperplasia) misclassified as DCIS. (e) A DCIS lesion misclassified as normal/benign. (f) IDC misclassified as DCIS.

Table 6.3: Confusion matrix of test set predictions

	BENIGN	DCIS	IDC
BENIGN	29	2	0
DCIS	4	12	4
IDC	0	2	11

normal/benign, DCIS, and IDC categories. CAS-CNN consists of two stages: in the first, we trained a CNN to learn cellular level features from small high-resolution patches and in the second, we stacked a fully convolutional network on top of this to allow for incorporation of global interdependence of structures to facilitate predictions in local regions. Our empirical evaluation demonstrates the efficacy of the proposed approach in incorporating more context to afford a high classification performance. CAS-CNN trained on large input patches outperforms the wide ResNet trained with input patches of size 224×224 by a large margin and consistently yields better results when trained with larger input patches.

Our system achieves an AUC of 0.962 for the binary classification of normal/benign slides from cancerous slides. This is remarkable, given the existence of 10 benign categories in the dataset, demonstrating the potential of our approach for pathology diagnostics. Based on the achieved performance on an independent test set, this system could be used to sieve out approximately 50% of obviously normal/benign slides on our dataset without missing any cancerous slides.

The performance of the system on the 3-class classification of WSIs was also very promising. An accuracy of 0.812 and a Kappa value of 0.700 were achieved. While discrimination of normal/benign slides from IDC slides was without any misclassification, errors in discriminating between normal/benign slides and DCIS slides, as well as DCIS and IDC slides were common. We postulate that the reason for these misclassifications is primarily because of the difficulty in discrimination of several benign categories such as usual ductal hyperplasia from DCIS which is also a source of subjective interpretation among pathologists. This could, in turn, be alleviated by obtaining more training data for these specific benign classes. The second reason could be the requirement of even larger receptive fields to enable discrimination of DCIS from invasive cancer. As seen in Table 6.1, the performance of CAS-CNN consistently improved with increasing patch size. However, this comes with increased computation time both during training and inference. One way to redress the problem could be the inclusion of additional downsampled patches with larger receptive fields as input to a multi-scale network[139–141] or using alternative architectures such

as U-net[142,143]. The final reason behind these errors lies in the fact that discrimination of certain DCIS patterns from IDC, purely based on H&E staining, can be complex. As such, pathologists may use additional staining such as myoepithelial markers to differentiate between DCIS and IDC lesions[144].

Although the current system learns to exhibit some hue and saturation invariance, specialized stain standardization techniques exist[43,56,96] and have been shown to greatly improve CAD system performance[34,145] by reducing the stain variations[58]. It is likely that standardizing the WSIs would also improve generalization of the performance of our network.

Although our primary aim was to facilitate pathology diagnostics by discriminating between different breast lesion categories, our system could serve as an important first step for the development of systems that aim at finding prognostic and predictive biomarkers within malignant lesions[146]. This will be one of our major directions for future work.

Deep neural networks for detection of breast cancer metastases

7

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM van der Laak, The CAMELYON16 Consortium;

Original title: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer

Accepted for publication in: Journal of the American Medical Association (JAMA), 2017

Abstract

The sentinel lymph node (SLN) procedure is pivotal for staging breast cancer. SLN tissue section evaluation by pathologists is laborious, has suboptimal accuracy, and can potentially be reinforced by automated computational methods. In this work, we assess the performance of automated deep learning systems at detecting metastases in hematoxylin and eosin stained tissue sections of lymph nodes of patients with breast cancer and compare it to pathologists in a diagnostic setting. In the setting of a competition, the results demonstrated that some deep learning algorithms achieved better diagnostic performance than a panel of 11 pathologists participating in a simulation exercise designed to mimic routine pathology workflow; performance was comparable to an expert pathologist in the absence of time constraints. Whether this approach has clinical utility will require evaluation in a clinical setting.

7.1 Introduction

Pathology diagnostics is at the doorstep of a digital revolution. Advances in slide scanning technology and cost reduction in digital storage capacity enable full digitalization of the microscopic evaluation of stained tissue sections ('digital pathology'). The advantages of these developments are many[7]: remote diagnostics, immediate availability of archival cases, easier consultations with expert pathologists, and perhaps the most significant benefit, the possibility for computerized or computer-aided diagnostics[8].

At the forefront of computerized diagnostics and machine learning research are 'deep learning' techniques, which enable computers to solve perceptual problems such as image recognition[31,147,148] and speech recognition[124,149–151]. Deep learning has already accomplished drastic improvements in several healthcare areas, including molecular activity prediction for drug discovery[152], predicting effects of non-coding variants[153], and understanding the genetic determinants of disease[154]. Notwithstanding significant breakthroughs, computerized clinical diagnostics in areas of medicine involving analysis of images (e.g. pathology, radiology, and ophthalmology) has yet to reach the level that can safely complement or even replace human observation. Gulshan et al.[127] and Esteva et al.[128] recently demonstrated the substantial potential of deep learning for diabetic retinopathy screening and automated skin lesion classification, respectively. Image analysis of pathology slides is also an important application of deep learning, but requires evaluation for diagnostic performance.

Breast cancer is a leading cause of cancer death in women worldwide. Microscopic assessment of sentinel lymph nodes (SLN) from breast cancer patients to determine the extent of cancer spread is an important part of breast cancer staging. The sensitivity of this subjective assessment is, however, far from optimal. A retrospective study in The Netherlands showed that pathology review by experts changed the nodal status (the N-classification in the TNM system[155]) in 24% of patients[156], while even presence of small clusters of tumor cells in the SLN are associated with worse prognosis and may require more aggressive treatment[157,158]. Furthermore, SLN assessment is a tedious and time-consuming component of pathologists' workload. In our earlier study[130], we found that all SLN slides containing tumor could be identified automatically while 40% of the slides without tumor could be excluded. This could result in a significant reduction in pathologists' workload.

Assessing the full potential of machine learning in digital pathology is difficult. Most approaches are evaluated on relatively small, single-center datasets, applying varying evaluation and reference standard methodologies. In many studies, a fair

and direct comparison of the newly proposed solution with the existing state-of-the-art methods and human performance is lacking. Consequently, it remains difficult to select promising methods for application in clinical practice.

The aim of the present study is to establish the state of the art of machine learning methods for the detection of metastases in SLN tissue sections and compare these to the performance of pathologists. To achieve this, we organized the CAMELYON16 grand challenge (CAncer MEtastases in LYmph nOdes challeNge) in conjunction with the IEEE International Symposium on Biomedical Imaging (ISBI). A grand challenge is a widely used, powerful means to compare varying approaches in a fair and reproducible way[159]. In CAMELYON16, research groups around the world were invited to produce an automated solution for breast cancer metastases detection in SLN, using the same data sets and were independently evaluated by the same, strict and predefined criteria. This article presents the results of the CAMELYON16 challenge and assesses the value of computerized digital pathology in a routine diagnostic setting by comparing algorithm performances to the results from a panel of pathologists.

7.2 Methods

7.2.1 Image datasets

To enable the development of diagnostic machine learning algorithms, we collected 399 whole-slide images of sentinel axillary lymph nodes (SLN) during the first half of 2015. SLNs were retrospectively sampled from 399 patients that underwent surgery for breast cancer at one of two hospitals in the Netherlands. The need for informed consent was waived by the institutional review board (research ethics committee of Radboudumc; file number 2016-2761). Whole-slide images where anonymized before making them available. All pathologists participating in this study were informed of, and agreed with the rationale and goals of this study. The participation of the pathologists was voluntarily and in accordance with the applicable Dutch rules concerning the review of research ethics committees and informed consent. In addition, the need to obtain informed consent from the panel of pathologists who participated in the study was waived by the research ethics committee.

To enable assessment of the performance of the algorithms for slides containing micro and macrometastases as well as for negative slides, stratified random sampling was performed on the basis of the original pathology reports.

Whole-slide images were acquired at two different centers – Radboud University Medical Center (RUMC), and University Medical Center Utrecht (UMCU) – using

two different scanners. RUMC whole-slide images were produced with a 3DHISTECH Pannoramic 250 Flash II digital slide scanner with a 20X objective lens (specimen level pixel size $0.243\mu m \times 0.243\mu m$). UMCU whole-slide images were produced using the Hamamatsu XR C12000 digital slide scanner with a 40X objective lens (specimen level pixel size $0.226\mu m \times 0.226\mu m$).

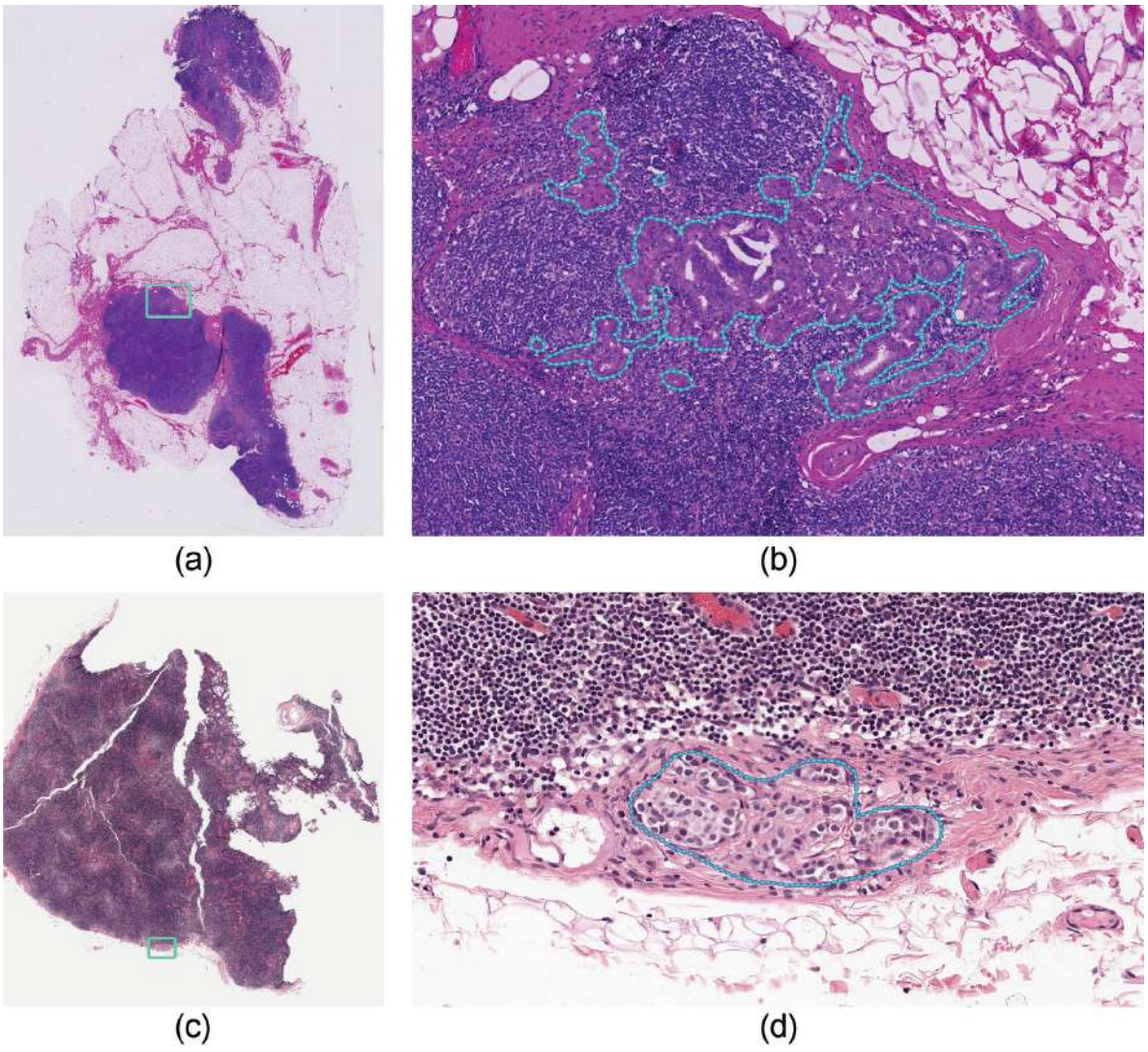

Figure 7.1: Two example annotated areas of WSIs taken from the CAMELYON16 dataset. (a) and (c) show overviews of two examples of WSIs. (b) and (d) are magnified images, corresponding to rectangle areas in *(a)* and *(c)*, with exhaustive annotation of metastatic regions.

7.2.2 Reference standard

All metastases present in the slides were annotated under the supervision of expert pathologists. The annotations were first manually drawn by two students (one for

each hospital) and then every slide was checked in detail by one of the two pathologists (Figure 7.1). In clinical practice pathologists may opt to use immunohistochemistry (IHC) to resolve diagnostic uncertainty. In this study, obvious metastases were annotated without the use of IHC whereas, for all difficult cases and all cases appearing negative on H&E, IHC (anti-cytokeratin mouse monoclonal antibody, clone CAM 5.2, BD Biociences, San Jose, USA) was used (Figure 7.2). This minimizes false negative interpretations. IHC is the most accurate method for metastasis evaluation and has little interpretation variability[160–162].

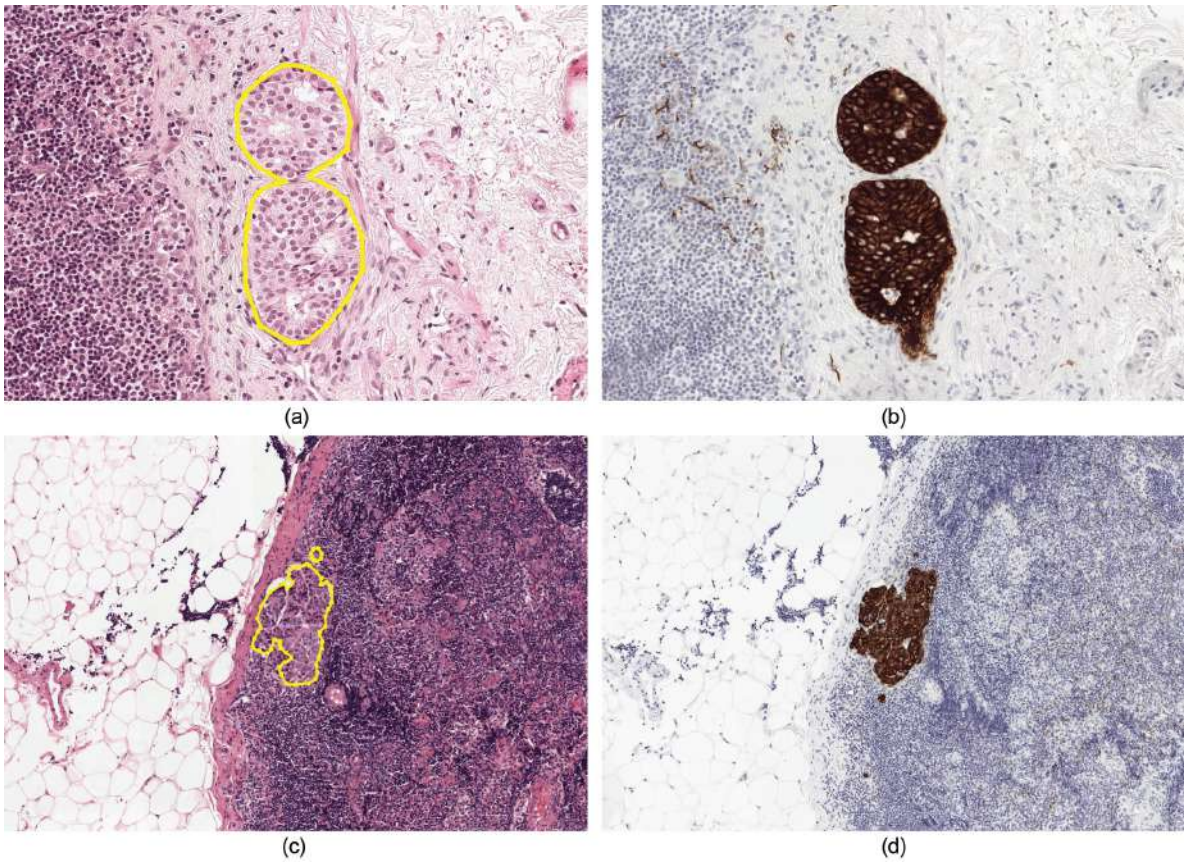

Figure 7.2: Side by side visualization of H&E and IHC staining for generating reference standard. (a) and (c) show two example annotations made for two H&E stained images. (b) and (d) show corresponding tissue areas in *(a)* and *(c)*, stained with IHC. Note that IHC was only used for generating the reference standard in our challenge. Neither of the pathologists in our observer study nor participants of the challenge had access to this data.

Pathologists differentiate between macrometastases (tumor cell cluster diameter$>$ $2mm$), micrometastases (tumor cell cluster diameter between 0.2 and $2mm$) and isolated tumor cells (ITC; solitary tumor cells or tumor cell clusters with diameter $\leq$ $0.2mm$ and/or less than 200 cells). The largest available metastasis determines the slide-based diagnosis. Because the clinical value of having only ITC in an SLN

is disputed, we did not include such slides in our study and also did not penalize missing ITCs in slides containing micro- or macrometastases. ITCs were, however, annotated in slides containing micro and macrometastases by the pathologists and included in the training WSIs. The set of images was randomly divided into a training ($n = 270$) and a test set ($n = 129$; details in Table 7.1). Both sets included slides with both micro and macrometastatic tumor foci as encountered in routine pathology practice.

Table 7.1: Number of WSIs per class in the training and testing sets.

dataset	hospital	macro	micro	IDC	non-IDC	nomral	total images
train	Radboudumc	35	35	54	16	100	170
	UMCUtrecht	26	14	30	10	60	100
test	Radboudumc	14	15	23	6	50	79
	UMCUtrecht	8	12	15	5	30	50

7.2.3 Coding challenge

In the first (training) stage of the CAMELYON16 challenge, participants were given access to 270 whole-slide images of digitally scanned tissue sections. Each SLN metastasis in these images was annotated enabling participants to build their algorithms. In the evaluation stage, the performance of the participants' algorithms was tested on a second set of whole-slide images lacking annotation of SLN metastases. The output of each algorithm was sent to the challenge organizers by the participants for independent evaluation. Each team was allowed to make a maximum of three submissions. Multiple submissions were only allowed if the methodology of the new submission was distinct.

7.2.4 Tasks and evaluation metrics

Two tasks were defined: 1. identification of individual metastases in WSIs, and 2. classification of every WSI as either containing or lacking SLN metastases. The tasks had different evaluation metrics and consequently resulted in two independent algorithm rankings. In task 1, algorithms were evaluated for their ability to identify specific metastatic foci in a whole-slide image. Challenge participants provided a list of potential metastasis locations with accompanying confidence scores in the range from 0 to 1. Algorithms were compared using a measure derived from the free-response receiver operator characteristic curve (FROC)[163]. The FROC curve shows

the lesion-level true positive fraction versus the mean number of false positive detections in metastasis-free slides only. The FROC true positive fraction score that ranked teams in the first task was defined as the mean true positive fraction at 6 predefined false positive rates: $1/4, 1/2, 1, 2, 4,$ and 8 FPs per WSI. Details on detection criteria for individual lesions can be found in the Appendix A.

Task 2 evaluated the algorithms' ability to discriminate between 49 slides with and 80 without SLN metastases. In this case, identification of specific foci within images was not required. Participants provided a confidence, using the same rating schema as for task 1, indicating the probability that each whole-slide image contained any evidence of SLN metastasis from breast cancer. The area under the receiver operating characteristic curve (AUC) was used to compare the performance of the algorithms.

7.2.5 Performance of human experts

To establish a baseline for human expert performance, two experiments were conducted using the 129 slides in the test set, corresponding to the tasks defined above. In the first experiment, one expert pathologist marked every single metastasis on a computer screen using high magnification. This task was performed without any time constraint. For comparison with the algorithms on task 2, the pathologist without time constraint indicated (during the same session) the locations of any (micro or macro) metastases per whole-slide image.

The setup without time constraint does not yield a fair measure of the accuracy of the routine diagnostic process. Preliminary experiments with four independent pathologists determined that two hours was a realistic amount of time for reviewing these 129 whole-slide images. We, therefore, asked 11 pathologists to independently assess the 129 glass slides in the test set in a simulation exercise designed to mimic routine diagnostic pathology workflow: a time limit of two hours was set but exceeding this limit was not penalized and every pathologist was allowed time to finish the entire set. The panel of the 11 pathologists (mean age 47.7 years; range $31-61$ years) included one resident pathologist (third year of residency) and ten practicing pathologists (average number of years practicing 16.4; range $0-30$; 0 is for one pathologist who just finished a 5 year residency program). Three of these pathologists had breast pathology as a special interest area.

The panel of 11 pathologists assessed the glass slides using a conventional light microscope and determined whether there was or was not any evidence of SLN metastasis in each image. This diagnostic task was identical to and used the same images as those evaluated by the algorithms in task 2. Pathologists indicated the level

of confidence in their interpretation for each slide using five levels: 'definitely normal', 'probably normal', equivocal', 'probably tumor', 'definitely tumor'. To obtain an empirical ROC curve, the threshold was varied to cover the entire range of possible ratings by the pathologists, and the sensitivity was plotted as a function of the false positive fraction (1 - specificity). To get estimates of sensitivity and specificity for each pathologist, the five levels of confidence were dichotomized by considering the confidence levels of 'definitely normal' and 'probably normal' as negative and all other levels as positive.

Statistical analysis

All statistical tests used in this study were two-sided considering a p-value < 0.05 as significant.

For task 1 and 2, confidence intervals of the FROC true positive fraction scores and areas under the ROC curve (AUC) were obtained using the percentile bootstrap method[164] for the algorithms, the panel of 11 pathologists, and the pathologist without time-constraint. The AUC values for the pathologists were calculated based on their provided five-point, ordinal scores.

For comparison of the AUC of the individual algorithms against the panel of 11 pathologists in task 2, we used multiple reader multiple case ROC (MRMC) analysis. The MRMC paradigm is frequently used for evaluating the performance of medical image interpretation and allows comparison of multiple readers reading the same cases while accounting for the different components of variance contributing to the interpretations[165,166]. Both the panel of readers and the algorithms as well as the cases were treated as random effects in this analysis. The panel of 11 pathologists represent the group of readers for modality one (diagnosing on glass slides) and an algorithm represents the reader for modality two (whole-slide images). Cases are the same set of slides/images seen by the panel and the algorithm. The AUC was the quantitative measure of performance in this analysis. The Dorfman-Berbaum-Metz significance testing, with Hillis improvements[167] was performed to test the null hypothesis that all effects are zero. The Bonferroni correction was used to adjust the p-values for multiple comparisons in the MRMC analysis (independent comparison of each of the 32 algorithms and the panel of pathologists).

Additionally, a permutation test was performed to assess whether there was a statistically significant difference between the area under the ROC curve of the pathologists detecting macro-metastases compared to micro-metastases[168]. This test was also repeated for comparing the performance of pathologists for different histotypes: infiltrating ductal cancer versus all other histotypes. As the controls (slides not containing metastases) were the same in both groups, the permutation was only per-

formed across the slides containing metastases. This test was performed for each individual pathologist and subsequently Bonferroni correction was applied to the obtained p-values.

No prior data was available for the performance of algorithms in this task. Therefore, no power analysis was used to predetermine the sample size.

iMRM 3.2 application developed by the Food and Drug Administration was used for MRMC analysis (available at https://github.com/DIDSR/iMRMC). An in-house developed script in Python 2.7 was used to obtain the percentile bootstrap CIs for the FROC and AUC scores. A custom script was written to perform the permutation tests and can be found at the same location.

7.3 Results

The expert pathologist without time constraint required approximately 30 hours for assessing 129 whole-slide images. She did not produce any false positives in task 1 (i.e. non-tumorous tissue indicated as metastasis) but failed to identify 27.6% of individual metastases (lesion-level true positive fraction of 72.4% (95% CI, 64.3%-80.4%)) that were manifest when immunohistochemical staining was performed. At the slide level (task 2), the pathologist achieved a sensitivity of 93.8% (95% CI, 86.9%-100.0%), a specificity of 98.7% (95% CI, 96.0%-100.0%), and an AUC of 0.966 (95% CI, 0.927-0.998). The 11 pathologists in the simulation exercise spent a median of 120 minutes for 129 slides (range, $72-180$ minutes). They achieved a mean sensitivity of 62.8% (95% CI, 58.9%-71.9%) with a mean specificity of 98.5% (95% CI, 97.9%-99.1%). The mean AUC was 0.810 (range, 0.738-0.884) ((Table 7.2 and Table 7.3 show results for individual pathologists). The ROC curves for each of the 11 pathologists and the pathologist assessing the slides without time constraint can be found in the online supplement.

7.3.1 Stratification according to metastasis size and primary tumor histotype in task 2

The pathologists' results were further analyzed in two subcategories: analysis according to metastasis size and primary tumor histotype (Table 7.2 and Table 7.3). Pathologist WTC achieved a better sensitivity and AUC for detecting macrometastases (sensitivity of 100% and AUC of 0.994 (95% CI, 0.977-1.0)) and metastases originating from invasive ductal carcinoma (IDC) (sensitivity of 97.0% (95% CI, 89.7%-100%) and AUC of 0.976 (95% CI, 0.932-1.0)) compared to micrometastases (sensitivity of 88.8% (95% CI, 75.0%-100%) and AUC of 0.943 (95% CI, 0.868-0.995)) and

Table 7.2: Classification results by the panel of 11 pathologists participating in the simulation exercise and the expert pathologist whiteout time constraint (WTC) on the CAMELYON16 test set for the whole-slide image classification task (task 2). We report sensitivity and specificity for different scenarios: 1) differentiating all tumor slides from normal slides, 2) differentiating slides with macrometastases from normal slides while excluding micrometastases, 3) differentiating slides with micrometastases from normal slides while excluding macrometastases, 4) differentiating slides with primary tumor histotype of IDC from normal slides while excluding the rarer primary tumor histotypes (non-IDC), and 5) differentiating slides with non-IDC primary histotypes from normal slides while excluding slides with primary tumor histotype of IDC.

Codename	All cases		Macrometastases		Micrometastases		IDC		Non-IDC	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Pathologist 1	0.612	1	0.954	1	0.333	1	0.647	1	0.533	1
Pathologist 2	0.510	0.987	0.909	0.987	0.185	0.987	0.588	0.987	0.333	0.987
Pathologist 3	0.632	1	0.954	1	0.370	1	0.735	1	0.4	1
Pathologist 4	0.653	0.987	0.954	0.987	0.407	0.987	0.705	0.987	0.533	0.987
Pathologist 5	0.755	0.987	1	0.987	0.555	0.987	0.764	0.987	0.733	0.987
Pathologist 6	0.571	0.975	0.818	0.975	0.370	0.975	0.676	0.975	0.333	0.975
Pathologist 7	0.469	0.975	0.863	0.975	0.148	0.975	0.529	0.975	0.333	0.975
Pathologist 8	0.632	0.975	0.954	0.975	0.370	0.975	0.705	0.975	0.466	0.975
Pathologist 9	0.571	0.987	0.909	0.987	0.296	0.987	0.617	0.987	0.466	0.987
Pathologist 10	0.734	0.962	0.954	0.962	0.555	0.962	0.794	0.962	0.6	0.962
Pathologist 11	0.775	1	0.954	1	0.629	1	0.850	1	0.6	1
mean pathologist	0.628	0.985	0.929	0.985	0.383	0.985	0.692	0.985	0.484	0.985
Pathologist WTC	0.938	0.987	1	0.987	0.888	0.9875	0.970	0.987	0.866	0.987

Table 7.3: Classification results by the panel of 11 pathologists participating in the simulation exercise and the expert pathologist without time constraint (WTC) on the CAMELYON16 test set for the whole-slide image classification task (task 2). We report classification AUC for different scenarios: 1) differentiating all tumor slides from normal slides, 2) differentiating slides with macrometastases from normal slides while excluding micrometastases, 3) differentiating slides with micrometastases from normal slides while excluding macrometastases, 4) differentiating slides with primary tumor histotype of IDC from normal slides while excluding the rarer primary tumor histotypes (non-IDC), and 5) differentiating slides with non-IDC primary histotypes from normal slides while excluding slides with primary tumor histotype of IDC. We used percentile bootstrapping to construct 95% confidence interval.

Codename	All cases		Macrometastases		Micrometastases		IDC		Non-IDC	
	AUC	95% CI	AUC	95% CI	AUC	95% CI	AUC	95% CI	AUC	95% CI
Pathologist 1	0.809	0.732 - 0.876	0.976	0.918 - 1.0	0.673	0.577 - 0.777	0.817	0.729 - 0.899	0.791	0.665 - 0.916
Pathologist 2	0.756	0.679 - 0.82	0.948	0.874 - 1.0	0.599	0.510 - 0.672	0.785	0.696 - 0.858	0.689	0.569 - 0.831
Pathologist 3	0.807	0.738 - 0.876	0.976	0.916 - 1.0	0.669	0.562 - 0.757	0.861	0.779 - 0.937	0.685	0.566 - 0.825
Pathologist 4	0.820	0.744 - 0.885	0.976	0.915 - 1.0	0.692	0.590 - 0.787	0.847	0.762 - 0.922	0.758	0.623 - 0.891
Pathologist 5	0.873	0.802 - 0.926	1.0	1.0 - 1.0	0.769	0.659 - 0.859	0.878	0.797 - 0.949	0.862	0.737 - 0.969
Pathologist 6	0.786	0.711 - 0.854	0.924	0.838 - 0.993	0.674	0.577 - 0.76	0.844	0.758 - 0.921	0.656	0.543 - 0.778
Pathologist 7	0.738	0.663 - 0.805	0.930	0.843 - 1.0	0.582	0.502 - 0.65	0.773	0.683 - 0.854	0.658	0.548 - 0.791
Pathologist 8	0.796	0.715 - 0.866	0.969	0.904 - 1.0	0.654	0.549 - 0.739	0.835	0.743 - 0.91	0.707	0.576 - 0.854
Pathologist 9	0.779	0.707 - 0.845	0.948	0.869 - 1.0	0.642	0.545 - 0.72	0.803	0.710 - 0.884	0.727	0.599 - 0.857
Pathologist 10	0.862	0.796 - 0.927	0.976	0.917 - 1.0	0.769	0.651 - 0.859	0.893	0.815 - 0.957	0.793	0.670 - 0.919
Pathologist 11	0.884	0.816 - 0.941	0.976	0.917 - 1.0	0.808	0.704 - 0.908	0.924	0.845 - 0.983	0.793	0.660 - 0.919
mean pathologist	0.810	0.750 - 0.869	0.964	0.930 - 0.997	0.685	0.619 - 0.746	0.842	0.775 - 0.907	0.738	0.630 - 0.846
Pathologist WTC	0.966	0.927 - 0.998	0.994	0.977 - 1.0	0.943	0.868 - 0.995	0.976	0.932 - 1.0	0.943	0.848 - 1.0

non-IDC cases (sensitivity of 86.6% (95% CI, 66.7%-100%) and AUC of 0.943 (95% CI, 0.848-1.0)), respectively (no statistically significant difference for comparison of AUCs, p=0.87 (Bonferroni corrected) for comparison of the performance for the detection of micro and macrometastases, and p>0.99 for comparison of the performance for the detection of IDC and non-IDC cases). For the 11 pathologists in the simulated routine diagnostic setting, the performance was significantly higher for detection of macrometastases (mean sensitivity of 92.9% (95% CI, 90.5%-95.8%) and AUC of 0.964 (95% CI, 0.952-0.976)) compared to micrometastases (sensitivity of 38.3% (95% CI, 32.6%-52.9%) and AUC of 0.685 (95% CI, 0.648-0.721)). We also observed that metastases originating from IDC (mean sensitivity of 69.2% (95% CI, 65.4%-77.4%) and AUC of 0.842 (95% CI, 0.817-0.866)) were more often detected compared to non-IDC cases (mean sensitivity of 48.4% (95% CI, 43.2%-59.7%) and AUC of 0.738 (95% CI, 0.701-0.776)) (but not significantly).

7.3.2 Algorithm performance

Between November 2015 and November 2016, 390 research teams signed up for the challenge. Twenty-three teams submitted 32 methods for evaluation by the closing date (details of the participating teams and their methods can be found in the online supplemental material of the thesis). The majority of entries (25/32) were based on deep convolutional neural networks. Besides deep learning, a variety of other approaches were attempted by CAMELYON16 participants. Different statistical and structural texture features were extracted (e.g. color SIFT features[169], local binary patterns[68], features based on gray-level co-occurrence matrix[97], etc.) combined with widely used supervised classifiers (e.g. support vector machines[170], random forest classifiers[121]). The performance and ranking of the top ten entries for the two leaderboards are shown in Table 7.4. Overall, deep learning-based methods performed significantly better than other methods: the top-19 entries in both tasks all used deep convolutional neural networks as the underlying methodology. Detailed method description for the participating teams can be found in the eMethods section of the online Supplement of the thesis.

Task1: Metastases identification

The results of metastases identification, as measured by the FROC sensitivity score, are presented in Table 7.5. The best algorithm, from team HMS & MIT (II), achieved an overall sensitivity score of 0.807 (95% CI, 0.732-0.889). The algorithm by team HMS & MGH (III) ranked 2nd in this leaderboard with an overall sensitivity score of 0.760 (95% CI, 0.692-0.857). Figure 7.3 presents the FROC curves for the top-five

Table 7.4: Results of the submitted algorithms on the CAMELYON16 test set for the metastasis identification and the whole-slide image classification tasks. Algorithms are sorted based on their performance on the whole-slide classification task. The percentile bootstrap method was used to construct 95% confidence intervals for FROC true positive fraction scores (FROC scores) and AUCs. The results of the significant test with MRMC analysis for the comparison of each individual algorithm with the panel of 11 pathologist are provided. The p-values were adjusted for multiple comparisons using the Bonferroni correction in which the p-values are multiplied by the number of comparisons (32; comparison of the 32 submitted algorithms with the panel of pathologists). See eMethods in the online supplementary material of the thesis for algorithms contact information and detailed description of each algorithm.

Codename	WSI classification		Lesion identification		Comparison to the panel of 11 pathologists	approach	Remarks
	AUC	95% CI	FROC score	95% CI			
HMS & MIT II	0.9935	0.983 - 0.999	0.807	0.732 - 0.889	$p < 0.001$	GoogLeNet	Ensemble of two networks, use of stain standardization, extensive data augmentation, hard negative mining
HMS & MGH III	0.9763	0.941 - 0.999	0.760	0.692 - 0.857	$p < 0.001$	ResNet	Fine-tuning pre-trained network, Fully convolutional network
HMS & MGH I	0.9643	0.928 - 0.989	0.596	0.578 - 0.734	$p < 0.001$	GoogLeNet	Fine-tuning pre-trained network
CULab III	0.9403	0.888 - 0.980	0.703	0.605 - 0.799	$p < 0.001$	VGG-16	Fine-tuning pre-trained network, Fully convolutional network
HMS & MIT I	0.9234	0.855 - 0.977	0.693	0.600 0.819	$p = 0.11$	GoogLeNet	Ensemble of two networks, Hard negative mining
ExB	0.9156	0.858 - 0.962	0.511	0.363 - 0.620	$p = 0.02$	ResNet	Varying class balance during training
CULab I	0.9087	0.851 - 0.954	0.544	0.467 - 0.629	$p = 0.04$	VGG-16	Fine-tuning pre-trained network
HMS & MGH II	0.9082	0.846 - 0.961	0.729	0.596 - 0.788	$p = 0.04$	ResNet	Fine-tuning pre-trained network
CULab II	0.9056	0.841 - 0.957	0.527	0.335 - 0.627	$p = 0.16$	VGG-Net & ResNet	Fine-tuning pre-trained network, Cascade a VGG-Net that operates on low magnification images and the ResNet model that refines the results.
DeepCare	0.8833	0.806 - 0.943	0.243	0.197 - 0.356	$p > 0.99$	GoogLeNet	Fine-tuning pre-trained network

Table 7.5: Results of the submitted algorithms on the CAMELYON16 test set for the lesion identification task. We report the overall FROC scores and the sensitivity at several values for the mean number of false positives per WSI (FPs/WSI). Note that the pathologist scoring without time constraint had an overall sensitivity of 72.4% without any false positives. This would translate to a final FROC sensitivity score of 0.724 for the second leaderboard of the challenge.

Codename	Final score	$\frac{1}{4}$ FPsWSI	$\frac{1}{2}$ FPs per WSI	1 FP per WSI	2 FPs per WSI	4 FPs per WSI	8 FPs per WSI
HMS & MIT II	0.807	0.773	0.778	0.813	0.827	0.827	0.827
HMS & MGH III	0.760	0.667	0.707	0.747	0.791	0.818	0.831
HMS & MGH II	0.729	0.729	0.729	0.729	0.729	0.729	0.729
CULab III	0.703	0.591	0.640	0.680	0.733	0.769	0.804
HMS & MIT I	0.693	0.596	0.649	0.693	0.738	0.742	0.742
HMS & MGH I	0.596	0.556	0.587	0.609	0.609	0.609	0.609
RadboudUMC	0.575	0.493	0.524	0.569	0.600	0.631	0.631
CULab I	0.544	0.404	0.471	0.493	0.582	0.631	0.684
CULab II	0.527	0.440	0.476	0.524	0.560	0.582	0.582
ExB	0.511	0.458	0.507	0.516	0.520	0.533	0.533

Table 7.6: Results of the submitted algorithms on the CAMELYON16 test set for the WSI classification task. We report classification AUC for different scenarios: considering all the normal slides and 1) all tumor slides, 2) slides with macrometastases, 3) slides with micrometastases, 4) slides with primary tumor histotype of IDC, 5) slides with the rarer primary tumor histotypes (non-IDC). We used bootstrapping to construct 95% confidence interval.

Codename	All cases		Macrometastases		Micrometastases		IDC		Non-IDC	
	AUC	95% CI	AUC	95% CI	AUC	95% CI	AUC	95% CI	AUC	95% CI
HMS & MIT II	0.9935	0.983 - 0.999	0.9905	0.973 - 1.0	0.9972	0.989 - 1.0	0.9926	0.979 - 1.0	0.9954	0.983 - 1.0
HMS & MGH III	0.9763	0.941 - 0.999	1.0	1.0 - 1.0	0.9569	0.893 0.999	0.9785	0.928 - 1.0	0.9712	0.920 - 1.0
HMS & MGH I	0.9643	0.928 - 0.989	0.9932	0.983 - 1.0	0.9407	0.876 - 0.987	0.9724	0.946 - 0.993	0.9458	0.857 - 0.997
CULab III	0.9403	0.888 - 0.980	0.9875	0.961 - 1.0	0.9019	0.812 - 0.962	0.9529	0.909 - 0.983	0.9117	0.785 - 0.991
HMS & MIT I	0.9234	0.855 - 0.977	0.9596	0.862 - 1.0	0.8939	0.794 - 0.971	0.9055	0.807 - 0.978	0.9642	0.915 - 0.996
ExB	0.9156	0.858 - 0.962	0.9948	0.985 - 1.0	0.8509	0.749 - 0.932	0.9276	0.855 - 0.981	0.8883	0.777 - 0.973
CULab I	0.9087	0.851 - 0.954	0.9966	0.989 - 1.0	0.8370	0.742 - 0.913	0.9290	0.868 - 0.974	0.8625	0.750 - 0.960
HMS & MGH II	0.9082	0.846 - 0.961	1.0	1.0 - 1.0	0.8333	0.738 - 0.917	0.9118	0.833 - 0.968	0.90	0.795 - 1.0
CULab II	0.9056	0.841 - 0.957	0.9926	0.972 - 1.0	0.8347	0.722 - 0.925	0.9311	0.852 - 0.983	0.8479	0.720 - 0.953
DeepCare	0.8833	0.806 - 0.943	0.9705	0.903 - 1.0	0.8123	0.704 - 0.895	0.8932	0.808 - 0.954	0.8608	0.756 - 0.973

performing systems in the first leaderboard (FROC curves of all methods in the online supplementary material). Figure 7.4 shows several examples of metastases in the test set of CAMELYON16 and the probability maps produced by the top-three ranked algorithms.

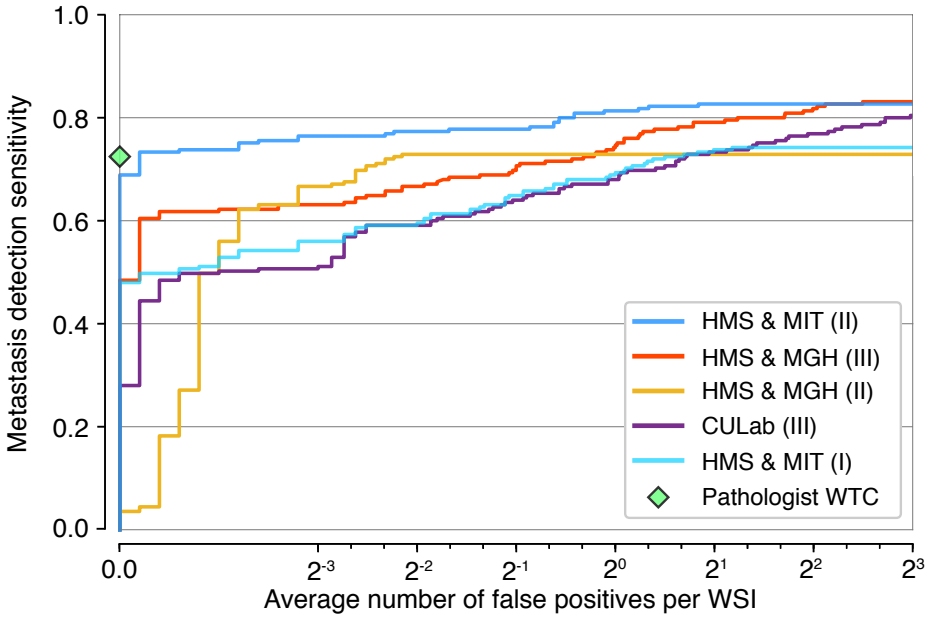

Figure 7.3: FROC curves of the top-five performing systems for the metastases identification task (task 1; measured on the 129 WSIs in the test set of which 49 contain metastatic regions). The range on the x-axis is linear between 0 to 2^{-3} and base-2 logarithmic scale between 2^{-3} and 2^3. Pathologist WTC refers to the pathologist who diagnosed the slides without time constraint. The pathologist did not produce any false positives and achieved a sensitivity of 0.724 for detecting and localizing metastatic regions.

Task 2: WSI classification

The results for all automated systems, sorted by their performance, are presented in Table 7.4. Figures 7.5a and 7.5b show the ROC curves of the top-five ranked teams along with the operating points of the pathologists. ROC curves for all the other methods can be found in the online supplementary material. All 32 algorithms were compared to the panel of pathologists using MRMC analysis (see Table 7.4).

The top-performing system by team HMS & MIT (II) was a GoogLeNet model[148] which outperformed all other CAMELYON16 submissions with an AUC of 0.993 (95% CI, 0.983-0.999). This AUC exceeded the mean performance of the 11 pathologists (mean AUC of 0.810 (95% CI, 0.750-0.869)) in our observer study ($p < 0.001$, calculated using MRMC analysis[166]). This AUC was comparable to that of the pathol-

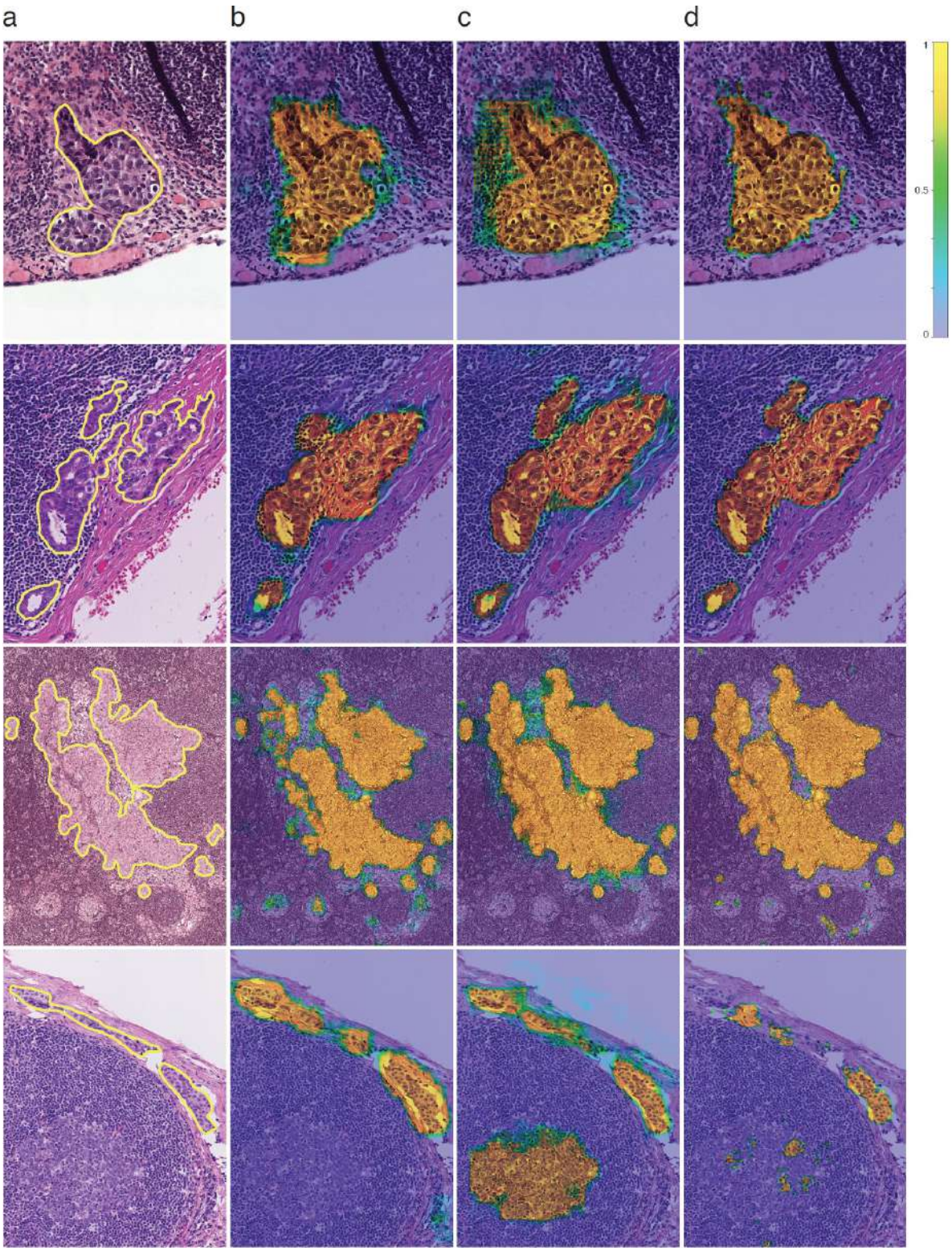

Figure 7.4: Example probability maps generated by the top-three performing systems. (a) Four annotated micrometastatic lesions in the test set of CAMELYON16. (b-d) Probability maps for teams HMS & MIT II, HMS & MGH III, and CULab III, respectively, overlaid on the original images.

ogist WTC (AUC = 0.966 (95% CI, 0.927-0.998)). Additionally, the operating points of all 11 pathologists were below the ROC curve of this method (Figure 7.5a). The ROC curves for the two leading algorithms, the pathologist WTC, and the mean ROC curve over the panel of 11 pathologists are shown in Figures 7.5c and 7.5d.

The second place algorithm by team HMS & MGH (III) used a fully convolutional ResNet-101[133] model. This system achieved an overall AUC of 0.976 (95% CI, 0.941-0.999), and yielded the highest AUC in detecting macrometastases (AUC 1.0). An earlier submission by this team, HMS & MGH (I), achieved an overall AUC of 0.965 (95% CI, 0.928-0.989) and ranked third. The fourth highest-ranked team was CULab (III) with a 16-layer VGG-net[135], followed by the first submission of the team HMS & MIT (I) with a 22-layer GoogLeNet. Overall, 7 of the 32 submitted algorithms had significantly higher AUCs than the panel of 11 pathologists (See Table 7.4 for individual p-values, calculated using MRMC analysis).

The top-ranking systems performed similarly to the best performing pathologists in detecting macrometastases (see Table 7.6). The performance of the algorithms in detecting micrometastases, however, was considerably more variable. Many of the top-ranked algorithms achieved better AUCs than the best pathologist in the panel of 11 (best pathologist AUC = 0.808 (95% CI, 0.704-0.908) versus best algorithm AUC = 0.997 (95% CI, 0.989-1.0)) in detecting micrometastases. The AUC of the two leading algorithms (AUC = 0.997 (95% CI, 0.989-1.0) and 0.957 (95% CI, 0.893-0.999), respectively) even surpassed that of the pathologist without time constraint (AUC = 0.9430 (95% CI, 0.868-0.995)).

With regard to the primary tumor histotype, the majority of the algorithms had higher AUCs for detecting IDC metastases than metastases of other types. The top-four performing algorithms achieved higher AUCs than the panel of 11 pathologists in detecting metastases of both IDC and non-IDC histotypes.

7.4 Discussion and conclusion

The CAMELYON16 challenge demonstrated that some deep learning algorithms were able to achieve better area under the ROC curve (AUC) than a panel of 11 pathologists for detection of lymph node metastases of breast cancer. To our knowledge, this is the first study that shows that interpretation of pathology images can be performed by deep learning based algorithms at an accuracy level that rivals human performance.

To obtain an upper limit on what level of performance could be achieved by visual assessment of H&E-stained tissue sections, a single expert pathologist exhaustively evaluated whole-slide images at high magnification and marked every single

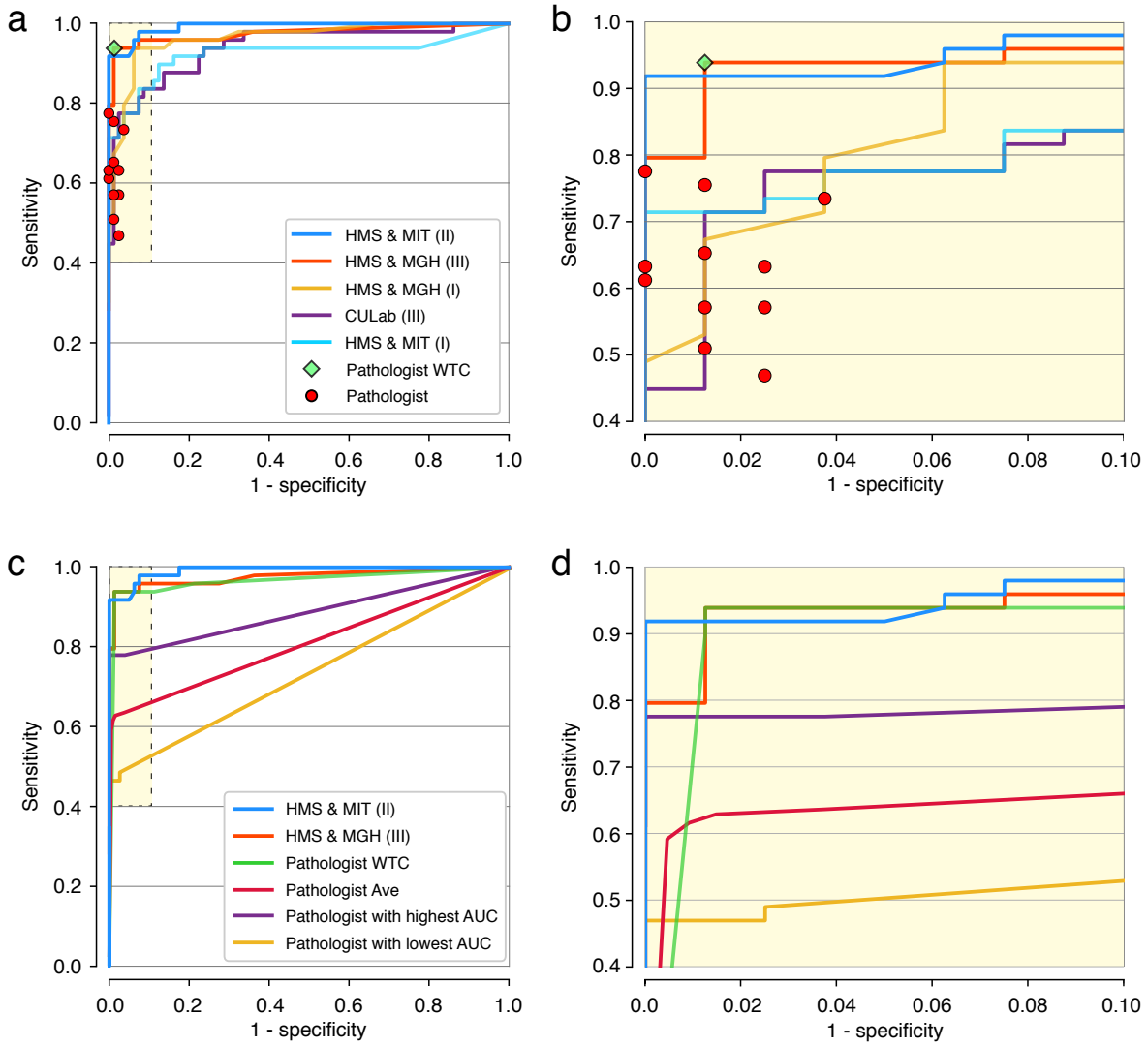

Figure 7.5: ROC curves of the top-performing systems (measured on the 129 WSIs in the test set of which 49 contain metastatic regions) for task 2. (a) ROC curves of the top-five performing systems and the operating points of the pathologists. Pathologist WTC refers to the operating point of the pathologist who diagnosed the slides without time constraint. All the pathologists scored WSIs using five levels of confidence: definitely normal, probably normal, equivocal, probably tumor, definitely tumor. For all pathologists, negative is defined as confidence levels 'definitely normal' and 'probably normal' and all others as positive. (b) More detailed view of the highlighted area in *(a)*. (c) Comparison of the ROC curves of the top-two performing systems, the mean ROC over the panel of 11 pathologists (denoted as Pathologist Avg.) and the pathologist WTC. (d) More detailed view of the highlighted area in *(c)*.

cluster of tumor cells. This took the pathologist 30 hours for 129 slides, which is infeasible in clinical practice. Although this pathologist was very good at differentiating metastases from false positives, 27.6% of metastases were missed compared

to the reference standard obtained with use of immunohistochemical staining to confirm the presence of tumor cells in cases where interpretation of slides was not clear-cut. This illustrates the relatively high probability of overlooking tumor cells in H&E-stained tissue sections. At the slide level, a high overall sensitivity and specificity for the expert pathologist, analyzing each case without time, constraint was observed.

To estimate the accuracy of pathologists in a routine diagnostic setting, 11 pathologists assessed the SLNs in a simulated exercise. The setting resembled diagnostic practice in The Netherlands, where use of IHC is mandatory for cases found negative on H&E reading. Compared to the expert pathologist interpreting the slides without time constraint, these pathologists were less accurate, especially on the slides which only contained micrometastases. Even the best performing pathologist on the panel missed over 37% of the cases with only micrometastases. Macrometastases were much less often missed. Specificity remained high, indicating that the task did not lead to a high rate of false positives.

The best algorithm achieved similar true positive fraction as the pathologist without time constraint when producing a mean of 1.25 false positive lesions in 100 whole-slide images and even performs better when allowing for slightly more false positives. On the slide level, the leading algorithms performed significantly better than the pathologists in a routine clinical setting.

All of the 32 algorithms submitted to CAMELYON16 used a discriminative learning approach to identify metastases in WSIs. The common denominator for the algorithms in the higher echelons of the ranking was that they used advanced convolutional neural networks. Algorithms based on manually engineered features had lower scores on average.

Despite the use of advanced convolutional neural network architectures, such as 16-layer VGG-Net[135], 22-layer GoogLeNet[148], and 101-layer ResNet[133], the ranking among teams using these techniques varied significantly, ranging from 1st to 29th (Appendix B illustrates some of the potential reasons for large variability in CNN performance). However, auxiliary strategies to improve system generalization and performance seemed more important. For example, team HMS & MIT improved their AUC in task 2 from 0.923 to 0.994 by adding a standardization technique[96] to help them deal with stain variations. Other strategies include exploiting invariances to augment training data (e.g. tissue specimens are rotation invariant), and addressing class imbalance (i.e. more normal tissue than metastases) by different training data sampling strategies (Appendix C contains further examples of properties that distinguish the best-performing methods).

Previous studies on diagnostic imaging tasks in which deep learning reached

human-level performance such as detection of diabetic retinopathy in retinal fundus photographs[127] used reference standard based on the consensus of human experts. This study, in comparison, generated reference standard using additional immunohistochemical staining, yielding an independent reference against which human pathologists could also be compared.

7.4.1 Limitations

This study has a number of limitations. The test dataset on which algorithms and pathologists were evaluated was enriched with cases containing metastases and, specifically, micro-metastases and thus is not directly comparable to the mix of cases pathologists encounter in clinical practice. Given the reality that most sentinel lymph nodes do not contain metastases, the dataset curation was needed to achieve a well-rounded representation of what is encountered in clinical practice without including an exorbitant number of slides. To validate the performance of machine learning algorithms, such as those developed in the Camelyon16 challenge, a prospective study is required. In addition, algorithms were specifically trained to discriminate between normal and cancerous tissue in the background of lymph node histological architecture, but might be unable to identify rare events such as co-occurring pathologies (e.g. lymphoma) or breast tissue. Finally, algorithm run-time was not included as a factor in the evaluation but might be relevant in, for example, frozen section analysis.

In this study, every pathologist was given a single H&E-stained slide per patient to determine the presence or absence of breast cancer metastasis. In a real clinical setting, multiple sections are evaluated for every lymph node, and typically multiple levels for each section are available. Also, in most hospitals pathologists request additional IHC staining in equivocal cases. Especially for slides containing only micrometastases, this is a relevant factor affecting diagnostic performance. The detection of other pathologies in the SLN (e.g. lymphoma) which is relevant in routine diagnostics was not included in the present study.

In addition, the simulation exercise invited pathologists to review 129 H&E stained slides in about two hours to determine the presence of macroscopic or microscopic SLN metastasis. Although feasible in the context of this simulation, this may not represent the work pace in other settings. Less time constraint on task completion may increase the accuracy of SLN diagnostic review. In addition, pathologists may rely on IHC staining and the knowledge that all negative H&E slides will undergo review.

7.4.2 Conclusions

In the setting of a challenge competition, some deep learning algorithms achieved better diagnostic performance than a panel of 11 pathologists participating in a simulation exercise designed to mimic routine pathology workflow; performance was comparable to an expert pathologist in the absence of time constraints. Whether this approach has clinical utility will require evaluation in a clinical setting.

Acknowledgment

The authors are grateful to the organizing committee of the 2016 IEEE International Symposium on Biomedical Imaging (ISBI) for hosting the workshop held as part of the study reported in this article. They also acknowledge their collaborators and financial support from their funding agencies. This work was partially funded by Stichting IT Projecten (STITPRO) Nijmegen, The Netherlands and by the Fonds Economische Structuurversterking (tEPIS/TRAIT project; LSH-FES Program 2009; DFES1029161 and FES1103JJT8U). The authors wish to acknowledge the financial support by the European Union FP7 funded VPH-PRISM project under grant agreement 601040. The authors would like to acknowledge the pathologists participating in the study. Ewout Schaafsma, Benno Kusters, Michiel vd Brand, Lucia Rijstenberg, and Michiel Simons from the department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands. Carla Wauters, Willem Vreuls, and Heidi Kusters from Canisius-Wilhelmina Hospital, Nijmegen, The Netherlands. Robert Jan van Suylen, Hans van der Linden, and Monique Koopmans from Jeroen Bosch Ziekenhuis, Den Bosch, The Netherlands. Gijs van Leeuwen, and Matthijs van Oosterhout from St. Antonius Ziekenhuis Nieuwegein and Peter van Zwam from Stichting PAMM, Eindhoven, The Netherlands. The authors declare no competing financial interests.

Supplementary material

The online supplementary material for this chapter can be downloaded from the following link: http://babakint.com

Appendix

A. CAMELYON16 evaluation metrics

In the lesion-based evaluation, a lesion was deemed to be identified if the location of the identified region was within the annotated reference standard lesion. If there were multiple findings for a single reference standard region, only the detection with the highest likelihood was considered while the lower likelihood findings were not considered false positives. All detections that were not within a specific distance ($\sim 75\mu m$) from the reference standard annotations were counted as false positives. In practice, there can be multiple small tumor regions that lie in the proximity of each other. Pathologists, however, consider all of these clusters as a single region. Therefore, it is important to consider them as a single lesion for the evaluation. We followed the guideline described by Cserni et al.[171] for merging these regions. Regions that were two or five cells apart ($\sim 75\mu m$) were considered as a single entity. Subsequently, we used the following steps to obtain the evaluation masks: (1) Applying distance transform on the inverse binary mask of reference standard, (2) Thresholding the distance transformed image ($T = 154$), (3) Labeling the connected components in the binary image. The resulting evaluation mask was a labeled image in which different tumor regions received different unique labels. This evaluation mask was used for the computation of the FROC curve.

B. Potential reasons for large variability in CNN performance

The modest performance of some of the CNN-based algorithms in many cases could be attributed to choosing a low magnification to process the slide, or selecting a very small patch size for training. Consequently, the system either lacks the detailed information present in the higher magnifications or loses the contextual information that could be captured by a larger patch size. Despite using the right magnification, patch size and state-of-the-art CNN architectures, achieving satisfactory results can be challenging. Training deep learning models can involve many hyperparameter settings (e.g. learning rate, regularization strength, mini-batch size, etc.). Successful and efficient training and debugging of large scale CNNs requires careful selection and adjustment on these hyperparameters, and finding out the relation between hyperparameters and validation errors.

C. Properties of the top-performing algorithms

We can summarize the main properties of the high-ranked teams based on 4 main characteristics: network architecture, patch-sampling strategy, preprocessing and data augmentation, and network ensemble.

One common property of the leading teams is that they all used very deep state-of-the-art CNN architectures such as GoogLeNet[148], VGG-Net[135], and ResNet[133]. The leading team, HMS & MIT (II), trained a 22-layer GoogLeNet model and enriched the training data by adding false positive findings produced by an initial model. By doing this, the network becomes more knowledgeable on recognizing the more difficult normal regions. The CNNs used in systems HMS & MGH (III), HMS & MGH (II), and CULab (III) were ResNet-101, GoogLeNet-22 and VGG-Net-16, respectively, all initialized by weights from pre-trained networks and fine-tuned with the challenge data. ResNet-101 was pre-trained on the MS-COCO dataset[172] and the other two models were pre-trained on the large scale 1000-class ImageNet dataset[125]. The high performance of these methods is in accordance with previous studies which have validated the efficacy of transfer learning strategies[173–176]. Some of the key factors contributing to the outstanding performance of the HMS & MIT (II) system were the use of the whole-slide image color standardizer (WSICS) algorithm[96] to normalize the appearance of WSIs, and the incorporation of a more rigorous data augmentation strategy including rotation, flipping, random cropping, and the addition of random offsets to each RGB color channel. The ResNet-101 model used in the system of HMS & MGH (III) used very large image patches of size 512512 that were more than double the input size of all the other systems used in this challenge. On top of that, the use of atrous convolution (dilated convolution) and spatial pyramid pooling[177] enabled the system to capture objects as well as image context at multiple scales.

Another factor contributing to the success of some of the top-ranking algorithms is the use of network ensembles. The winning team used an ensemble of a network trained on standardized WSIs and a network trained on original WSIs to report the probabilities for each finding. The first submission of this team, HMS & MIT (I), ranking fifth, used an ensemble of two networks (networks trained before and after hard-negative mining).

To generate a slide based score for the second task, the majority of the teams assigned the maximum probability among the detected lesions in the WSI as the confidence score for that slide. Prior to this assignment, they mostly removed small areas of positive findings, and/or applied Gaussian/median filtering. Although this approach worked well for many of the teams, including CU-Lab (III) and ExB research

that were ranked fourth and sixth in the image classification task, it may not take into account metastases characteristics (e.g. slides containing multiple high-score findings or slides containing larger metastases could have increased chance of containing metastases). In contrast, the systems HMS & MIT (I & II) used a random forest classifier employing a variety of geometrical and morphological features extracted from each probability map. Details of these features can be found in eMethods. The use of a learning-based algorithm to produce a confidence score from a WSI probability map is likely the centerpiece of this algorithm that makes it the top-performing system for the first task.

Finally, one interesting property of the top-performing system HMS & MIT (II) in the lesion identification task is that it uses the output of the discriminative classifier that produces a slide-based confidence score, to weigh the score of each finding in the second task. This top-down analysis reduces the number of false-positives, particularly in normal slides.

Using deep learning to identify and classify tumor-associated stroma

8

Babak Ehteshami Bejnordi, Maeve Mullooly, Ruth M. Pfeiffer, Shaoqi Fan, Pamela M. Vacek, Donald L. Weaver, Sally Herschorn, Louise A. Brinton, Bram van Ginneken, Nico Karssemeijer, Andrew H. Beck, Gretchen L. Gierach, Jeroen van der Laak, Mark E. Sherman

Original title: Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies

Submitted to: Under review

Abstract

The breast stromal microenvironment is a pivotal factor in breast cancer development, growth and metastases. Although pathologists often detect morphological changes in stroma by light microscopy, visual classification of such changes is subjective and non-quantitative, limiting its diagnostic utility. To gain insights into stromal changes associated with breast cancer, we applied automated machine learning techniques to digital images of 2,387 hematoxylin and eosin stained tissue sections of benign and malignant image-guided breast biopsies performed to investigate mammographic abnormalities among 882 women, ages 40-65, that were enrolled in the Breast Radiology Evaluation and Study of Tissues (BREAST) Stamp Project. Using deep convolutional neural networks, we trained an algorithm to discriminate between stroma surrounding invasive cancer and stroma from benign biopsies. In test sets (928 whole-slide images from 330 patients), this algorithm could distinguish biopsies diagnosed as invasive cancer from benign biopsies solely based on the stromal characteristics (area under the receiver operator characteristics (ROC) curve, AUC=0.962). Furthermore, without being trained specifically using DCIS as an outcome, the algorithm detected tumor-associated stroma in greater amounts and at larger distances from grade 3 versus grade 1 DCIS. Collectively, these results suggest that algorithms based on deep convolutional neural networks that evaluate only stroma may prove useful to classify breast biopsies and aid in understanding and evaluating the biology of breast lesions.

8.1 Introduction

The diagnostic classification of benign breast diseases, putative breast cancer precursors, and breast cancer is based largely on the histopathological appearance and molecular characteristics of epithelial cells[113]. Although the appearance of breast stroma contributes to pathologists' diagnostic impressions, including recognition of invasion, these subjective assessments have not been formally classified. Given that the tumor microenvironment is important in tumor growth, angiogenesis, and metastasis[178–180], and that stromal-epithelial interactions[181–184] contribute to progression of ductal carcinoma in-situ (DCIS) to invasive breast cancer, we hypothesize that morphological analysis of stroma could have importance in understanding breast carcinogenesis and diagnosis. This view is supported by evidence that the transition from DCIS to invasion is characterized by greater changes in gene expression of stromal cells than in epithelial tumor cells[185,186].

Apart from evaluation of lymphoid infiltrates, which are a diagnostic feature of medullary carcinoma and can be graded[187], stromal alterations are often subtle and difficult to characterize and quantify by light microscopy alone. Emerging data suggest that automated pattern recognition systems could be used to characterize stromal changes. For example, in a computer-generated automated analysis of routinely prepared hematoxylin and eosin (H&E) stained breast cancer tissue sections, Beck et al.[80] reported that stromal features were associated with breast cancer survival, and were more predictive of prognosis than epithelial features. Development of an automated computerized tool to identify and characterize tumor-associated stroma could have utility in pathological diagnosis with respect to evaluating tumor margins and cancer field effects or in predicting the potential of DCIS to progress to invasion, if occult neoplastic cells persist after treatment.

Development of robust computerized algorithms for discriminating patterns of normal stroma and tumor-associated stroma in histopathology images is a complex task, partly because validated morphologic criteria for distinguishing tumor-associated stroma are undefined. Machine learning approaches, and more specifically deep learning algorithms, could prove very suitable for accomplishing this objective as they are capable of learning the most discriminative features directly from a large set of classified diagnostic images, and therefore, do not require pre-defined morphologic criteria[36,123]. Thus, the aims of the present study were: 1) to generate a deep learning algorithm that can identify and characterize tumor-associated stromal alterations in H&E stained sections of breast biopsies; and 2) to assess stromal characteristics in DCIS in relation to grade, which may represent a proxy for risk of invasion.

8.2 Material and methods

8.2.1 Case selection

This analysis included 882 women, ages 40-65 years, referred for diagnostic image-guided breast biopsies (including ultrasound-guided needle core biopsy and stereotactic vacuum-assisted biopsy), who participated in the Breast Radiology Evaluation and Study of Tissues (BREAST) Stamp Project[188] undertaken between 2007 and 2010 at the University of Vermont Larner College of Medicine and the University of Vermont Medical Center. Women provided informed consent, which included access to medical records, self-reported breast cancer risk information, blood and saliva donations, access to radiological images and pathological tissues for research and follow-up. The study was approved by appropriate ethics review boards at the University of Vermont and at the National Cancer Institute (National Institutes of Health).

Breast biopsies were performed as ultrasound-guided core needle biopsies (14-gauge) or as stereotactically-guided vacuum-assisted biopsies (9-gauge) that were routinely fixed in formalin, prepared as paraffin-embedded tissue sections, and stained with H&E for diagnosis. For study purposes, biopsies were classified as non proliferative benign breast disease (BBD), proliferative BBD without atypia, atypical hyperplasia, ductal or lobular carcinoma in-situ or invasive carcinoma[189]. When biopsies included multiple tissue blocks, reflecting target and surrounding non-target tissues, we attempted to collect sections from both types of blocks, yielding a total of 2,387 total H&E stained sections that were scanned at 20X (Aperio, ScanScope CS or Hamamatsu) as digital images (resulting specimen level pixel size $0.445\mu m \times 0.445\mu m$).

8.2.2 Reference standard

Manual annotation of several tissue structures was made to train our deep learning algorithms. Different breast tissue components such as stroma, epithelium, and fat were annotated. In addition, stromal regions adjacent to invasive cancer were manually demarcated in the whole-slide images (WSIs) under the supervision of a pathologist, regardless of their visual appearance. Examples of normal stroma were also annotated in WSIs of benign biopsies (without DCIS or invasive breast cancer diagnosis). To analyze the pattern of stroma surrounding DCIS lesions, the WSIs containing only DCIS, and WSIs containing DCIS with concurrent invasive cancer were annotated by a pathologist (MES). For each case, a subset of ducts containing DCIS lesions was annotated on WSIs with point annotations in the center of the lesion and graded using standard criteria based on nuclear size and appearance, mitoses and detection of necrosis[190]. DCIS lesions in slides with concurrent invasive

cancer were annotated if they were peripheral to the invasive component and its associated stroma.

8.2.3 Deep learning algorithms

Deep learning is a subfield of machine learning, where very general algorithms learn features directly from data for prediction and classification. Our WSI classification system is based on multiple deep convolutional neural networks (CNN). To enable assessment of unbiased performance of our algorithm, the dataset was randomly split into a training set containing 62% of the WSIs (1459 WSI from 552 patients) and a testing set with the remaining slides (928 WSI from 330 patients; Table 8.1).

Table 8.1: Summary of number of whole-slide images (WSIs) from image-guided diagnostic breast biopsies and their diagnoses used for the development and validation of the proposed system.

Diagnosis	Entire dataset			Training dataset			Testing dataset		
	# Patient	# WSI	%	# Patient	# WSI	%	# Patient	# WSI	%
Benign	321	675	36.4	209	437	37.9	112	238	33.9
Proliferative	312	937	35.4	209	608	37.9	103	329	31.3
Proliferative with atypia	57	212	6.5	42	171	7.6	15	41	4.5
DCIS	58	222	6.6	–	–	–	58	222	17.6
LCIS	10	29	1.1	7	21	1.2	3	8	0.9
Invasive breast cancer	124	312	14.0	85	222	15.4	39	90	11.8
Total	882	2387	100	552	1459	100	330	928	100

Briefly, using representative input patches from the annotated areas, a model denoted '*CNN I*' was trained using the approach we previously described[191], to classify the WSI into three major tissue components: fat, stroma, and epithelium. Next, a second model '*CNN II*' was trained operating on stromal regions recognized by *CNN I*. *CNN II* assigned a score indicating the probability for the input stroma image to be classified as cancer-associated stroma. This model was trained using manually identified regions of stroma adjacent to invasive cancer and stroma in WSIs not containing tumor. Examples of DCIS-associated stroma were not used in the training phase. To classify WSIs into normal/benign vs invasive cancer, a third model '*CNN III*' was constructed and composed of a small CNN stacked on top of *CNN II*. *CNN III* was trained to generate a score for the entire WSI indicating the probability that the slide contained invasive cancer. More details about this network are described below. Figure 8.1 shows an overview of the entire classification system.

CNN I and *CNN II* in this study possess a VGG-Net-like architecture[135]. VGG-Net is a neural network architecture developed by Oxford's Visual Geometry Group (VGG), which won the 2014 ImageNet Large Scale Visual Recognition Challenge

2014, for the object localization task[125]. Details of our network configuration and training procedures are presented in the Supplementary material sections (at the end of this chapter): 'Convolutional neural network architecture', 'Preprocessing of WSIs and ground truth ROIs' and 'Training procedure'.

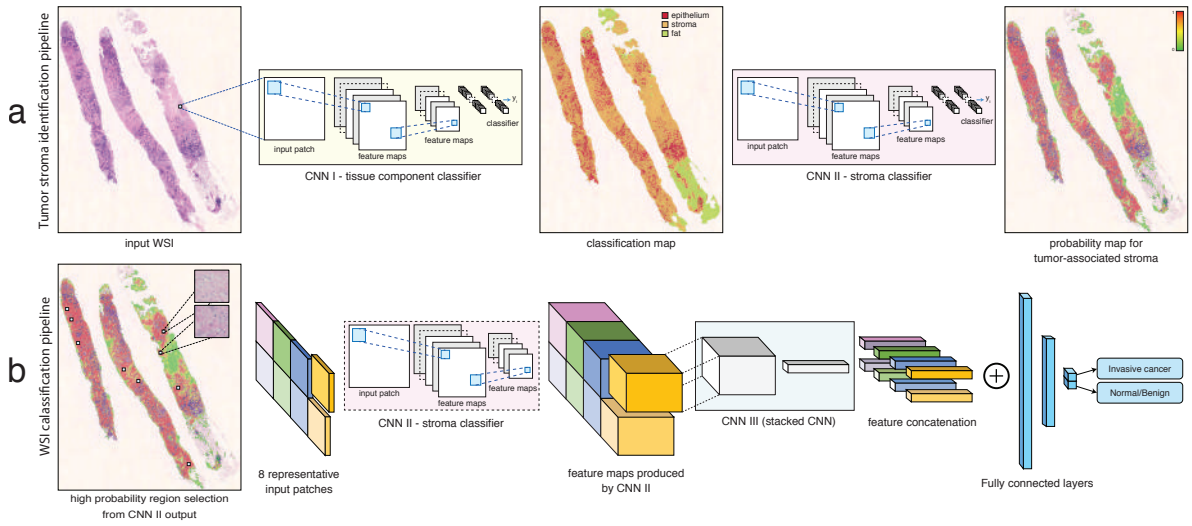

Figure 8.1: Overview of the system used for tumor-associated stroma identification and the system used for classification of the whole-slide image (WSI) into normal/benign or invasive cancer. (a) The top row shows the application of two convolutional neural networks (CNNs) for identifying regions of tumor-associated stroma in the WSI. *CNN I* classifies the tissue into epithelium, stroma, and fat. *CNN II* operates on stromal regions detected by *CNN I* and is trained to identify regions of tumor-associated stroma. (b) To classify the WSI into invasive cancer vs benign/normal, *CNN III* (stacked CNN) takes as input eight patches with high probabilities for tumor-associated stroma identified by *CNN II* and classifies the WSI by stacking a set of convolutional and fully connected layers on top of *CNN II* (see Table 8.2 for details).

8.2.4 Description of *CNN III* for the classification of WSIs into normal/benign vs invasive cancer

To illustrate the potential of stroma characterization, *CNN III* was constructed to identify cancerous biopsies based on the output of *CNN II* only. The output feature map of the penultimate layer of *CNN II* (the hidden layer whose output is fed to the final classification layer) is a compact representation of the input stromal image. *CNN III* takes as input the feature maps from *CNN II* for eight non-overlapping stromal tissue regions (size $152.9\mu m \times 152.9\mu m$), which were identified by *CNN II* as harboring the strongest tumor-associated alterations and predicted the WSI diagnosis (normal/benign vs invasive breast cancer; Figure 8.1b). Details of the procedure

for selecting those regions can be found in Supplementary section 'Selection of candidate stromal regions as input to *CNN III*'.

To generate the final score for each slide representing the probability of being a cancerous biopsy, we used an ensemble of two networks comprising *CNN III* and a modified version of *CNN III* without the last two fully connected layers. The average probability of the two networks was taken as the final score.

We additionally compared the performance of *CNN III*, developed in the present study, with our recently published approach for WSI classification[191]. Our previous system derived a total of 71 features from the outputs of both *CNN I* and *CNN II* and used these as input for a random forests classifier[121]. These features include the global tissue amount for epithelium, stroma, and fat as well as morphological features of epithelial areas and the spatial distribution of epithelial areas in the WSI derived from two region adjacency graphs: Delaunay triangulation and area-Voronoi diagram[192].

8.2.5 Experiments

1) Classification of breast tissue WSIs as invasive carcinoma versus BBD: The training data (1459 WSI from 552 patients) used for this classification task was further divided into two sets, a preliminary set to define parameters, and a second (validation) set (comprising 10% of slides) that was used to perform final model selection and hyper-parameter optimization. The performance of our model was evaluated on the independent test set (928 WSI from 330 patients) described previously above.

2) Analysis of DCIS-associated stroma: In this experiment, we analyzed the stromal patterns surrounding DCIS lesions on breast cancer slides. The DCIS-associated stroma was analyzed using *CNN II* which was trained to discriminate between normal and tumor-associated stroma. We first classified all the stromal pixels adjacent to annotated DCIS lesions using *CNN II*. Subsequently, we extracted two measures to quantify DCIS-associated stroma. These measures are the mean and standard deviation of all tumor-associated stroma probabilities for the pixels surrounding DCIS lesions. They were computed for a range of distances from the lesion's margin. This analysis was performed independently on test slides with DCIS lesions only and with test slides containing DCIS accompanied by invasive cancer.

8.2.6 Statistical analysis

The area under the receiver operator characteristic (ROC) curve (AUC) was used to evaluate the performance of the system in discriminating between invasive carcinoma and BBD biopsies. The ROC plots the sensitivity versus the false positive

fraction (1 - specificity). 95% confidence intervals for the ROC curves were obtained using the percentile bootstrap method[164]. The significance test for comparing two correlated ROC curves, when comparing the performance of the proposed system for classification of WSIs with our previously described system, was done using the bootstrap method in R package "pROC"[193]. This method is based on the approach described by Hanley and McNeil[70] that takes into account the correlation that is induced by the paired nature of the data.

The one-way analysis of variance (ANOVA) and the Tukey post hoc test were used to compare the computed stromal measures described in results section "Analysis of DCIS-associated stroma" for the patients with different DCIS grades. A p-value$<$0.05 was considered significant. All analyses were two-tailed.

8.2.7 Results

Classification of breast tissue WSIs as invasive carcinoma versus BBD

CNN I, subdividing WSIs into regions consisting of epithelium, stroma, and fat achieved a pixel-level 3-class classification accuracy of 95.5% compared to reference standard, computed on a balanced subset of annotated pixels in the independent test set. Representative examples of tissue classification results are shown in Figure 8.2. *CNN II* used for classifying stroma into normal stroma and tumor-associated stroma achieved a binary classification accuracy of 92.0% compared to reference standard, computed on a balanced subset of annotated pixels in the independent test set. Figure 8.2 also shows the output probability map for a normal slide (Figure 8.2d-f) and a slide containing invasive cancer (Figure 8.2a-c).

Figure 8.3 shows the ROC curves for the WSI classification of invasive cancer vs. non-cancer using our proposed system and our previously published method[191]. Our newly developed CNN model achieved an AUC of 0.962 (95% CI 0.936-0.983), which was slightly higher (but not statistically significantly, $p = 0.48$) than our previously described approach[191], which achieved an AUC of 0.948 (95% CI 0.915-0.977).

Analysis of DCIS-associated stroma

Figure 8.4 shows representative DCIS lesions with different histological grades and their corresponding probability maps for tumor-associated stroma. Figure 8.5 shows examples of stroma patches for different grades of DCIS harboring tumor-associated alterations. Boxplots for the mean and standard deviations of DCIS-associated stroma probabilities for the pixels $< 175\mu m$ from the DCIS margin are shown in Figure 8.6a and 8.6b. In Figure 8.6a and 8.6b, each point is a single DCIS lesion. Overall, the

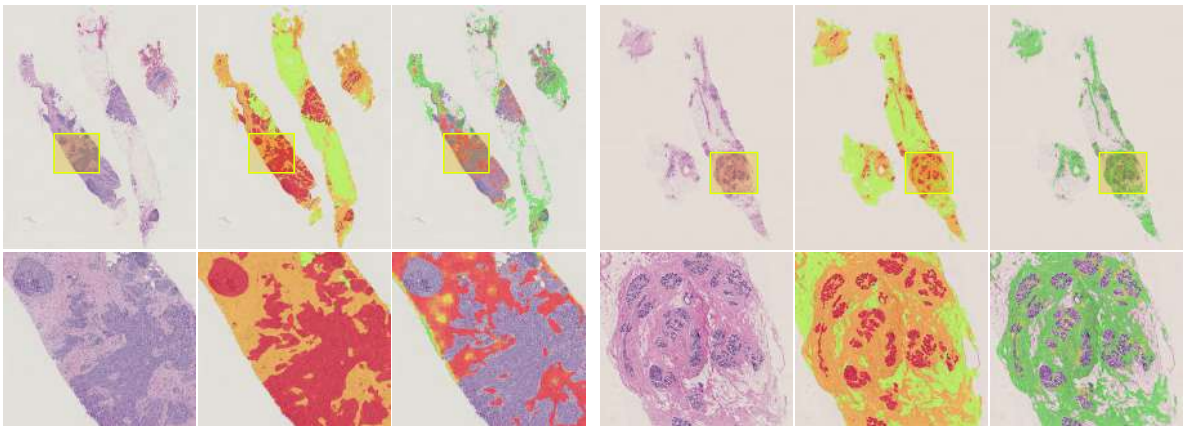

Figure 8.2: Example results for the epithelium, stroma, and fat classification and tumor-stroma identification. (a) shows a representative WSI containing invasive cancer. (b) The classification map of epithelium (red), stroma (orange), and fat (green) for the image *(a)*, produced by *CNN I*. (c) The map produced by *CNN II* showing the tumor-stroma probability for the image *(a)*, where green, yellow, and red represent low, medium and high probabilities, respectively. (d) A WSI with benign epithelium, mostly terminal duct lobular units (TDLUs). (e) and (f) show the results for the epithelium, stroma, and fat classification and tumor-stroma identification for the image *(d)*, respectively.

amount of tumor-associated stroma increased with increasing lesion grade. In addition, it was observed that DCIS lesions in slides without an invasive component presented less tumor-associated stroma compared to DCIS lesions in slides demonstrating invasive cancer. Figures 8.6c and 8.6d show similar boxplots at the patient level where the average tumor stroma probability is shown as computed by taking the mean of the scores for all DCIS foci that had the highest DCIS grade for the patient. A statistically significant difference was observed for the patient level means and standard deviations among the patients with different DCIS grades ($p = 0.023$ and 0.005, respectively; One-way ANOVA). For slides containing DCIS only, average tumor stromal probabilities were higher for higher grade DCIS lesions, but this relationship was not evident for DCIS associated with invasive cancer. Values for tumor stromal probabilities showed greater variability for higher grade lesions. After multiple comparisons adjustment, the mean DCIS-associated stroma probabilities (Figure 8.6c) were significantly different between DCIS grades 1 and 3 ($p = 0.028$). Mean differences between grades 2 and 3 and grades 1 and 2 were not statistically significant ($p = 0.217$ and $p = 0.329$). For the standard deviation of DCIS-associated stroma probabilities (Figure 8.6d), we observed statistically significant differences between grades 1 and 3 as well as grades 2 and 3 ($p = 0.021$ and $p = 0.028$). No

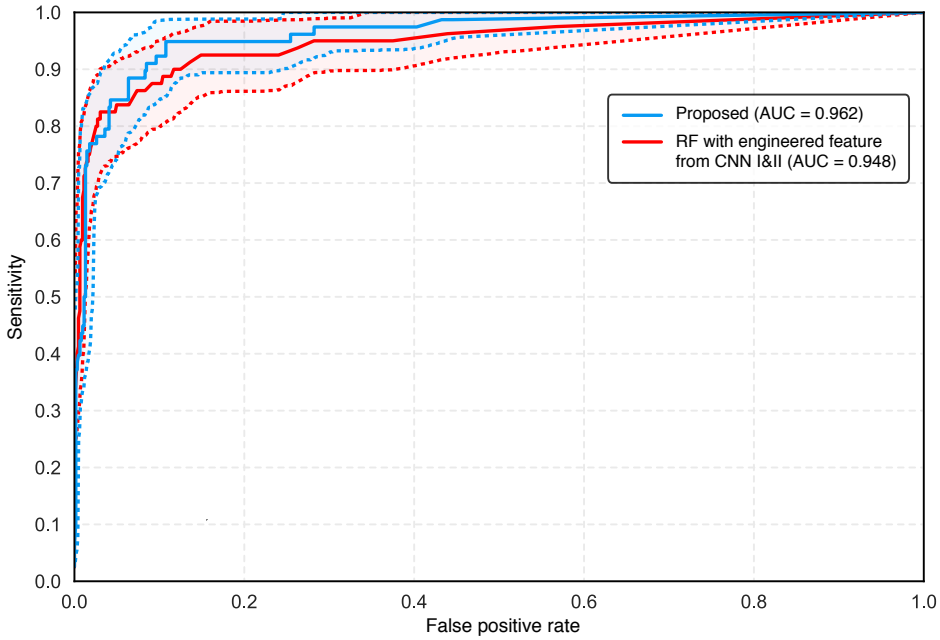

Figure 8.3: ROC curves with 95% confidence interval of the proposed system and the previously described approach[191] for the WSI classification task of identifying invasive cancer vs. benign breast biopsies. 95% confidence intervals for the ROC curves were obtained using the percentile bootstrap method[164].

statistical significance in standard deviation was observed between grades 1 and 2 ($p = 0.619$).

Figures 8.7a and 8.7b show the mean and standard deviation of DCIS-associated stroma for rims of stroma at different distances from the DCIS perimeter. As the distance from the DCIS margin increases, the mean of the DCIS-associated stroma probabilities decreases, but only slightly, and curves for different grades of DCIS remain parallel up to $500\mu m$ from lesion periphery.

8.3 Discussion and conclusion

In this study, we developed a state-of-the-art deep CNN for distinguishing BBD from invasive breast cancer based on the identification and characterization of tumor-associated stromal alterations. In an independent test set, classification of breast biopsies as benign or malignant based solely on CNN analysis of stroma achieved an impressive AUC=0.962, consistent with highly accurate discrimination. Without training the CNN on DCIS lesions, we subsequently assessed whether tumor-associated stroma could be identified in tissues surrounding DCIS and whether its extent varied with clinically important pathologic features. We detected greater amounts of tumor-associated stroma in grade 3 versus grade 1 DCIS and also found

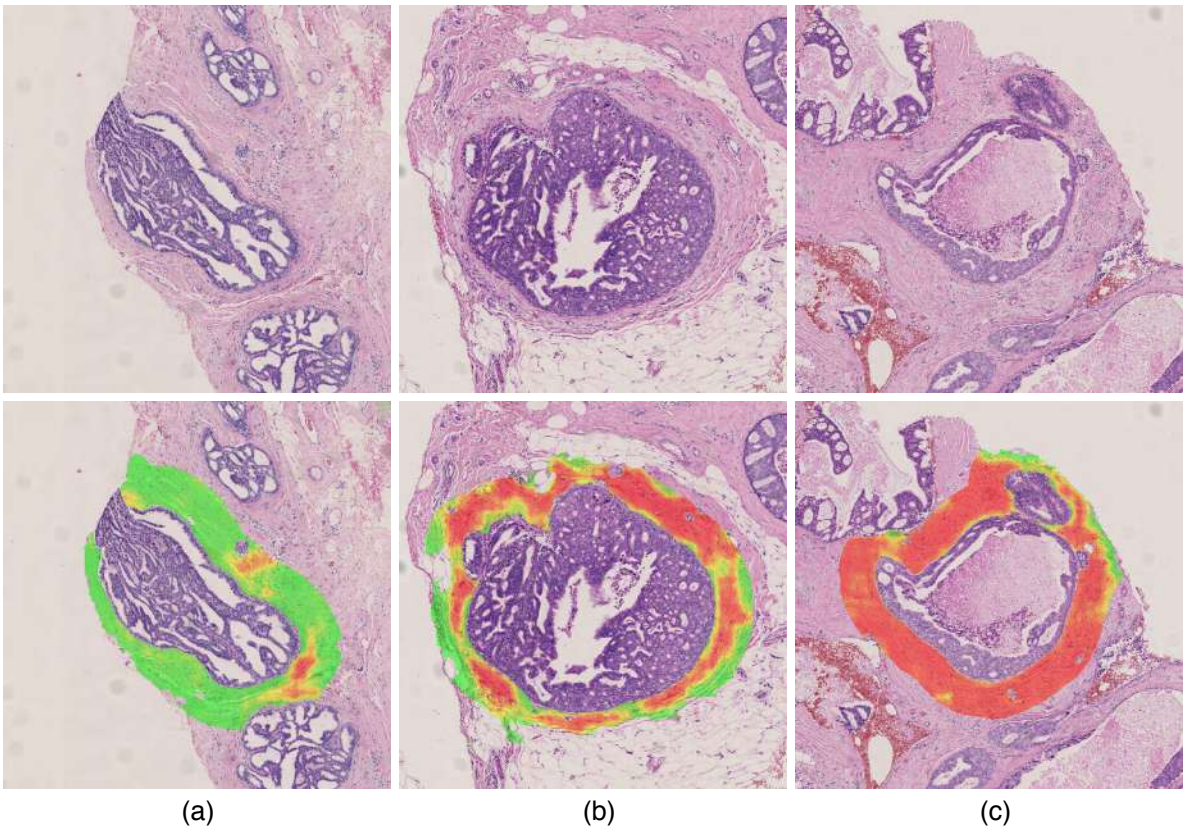

Figure 8.4: Example DCIS lesions with different histological grades and their corresponding probability maps generated by *CNN II* for identification of regions of tumor-associated stroma. The DCIS lesions shown in (a), (b), and (c) have histological grades of 1, 2 and 3 (with necrosis), respectively. The tumor-stroma probability maps (green, yellow, and red represent low, medium and high probabilities) overlaid on the original images are shown for the pixels with distances below $175\mu m$ from the DCIS margin.

that DCIS associated with an invasive component generally possessed higher amounts of tumor-associated stroma compared to slides containing DCIS only. Thus, our work provides support for including morphological analysis of breast stroma in studies aiming to understand risk of DCIS progressing to invasion and in defining the biology of invasive breast cancer.

To date, most previous work[86,87,90,100,114,194] using automated image analysis approaches to detect and classify breast cancer in histological images involved assessment of the morphology and arrangement of epithelial structures (e.g. nuclei, ducts). Generally, the aim of this research was to objectify, standardize and quantify features that are already appreciated as important by pathologists. Although subjective evaluation of stroma may provide cues that pathologists use in the histopathologic diagnosis of breast lesions, stroma is difficult to assess microscopically, and formal

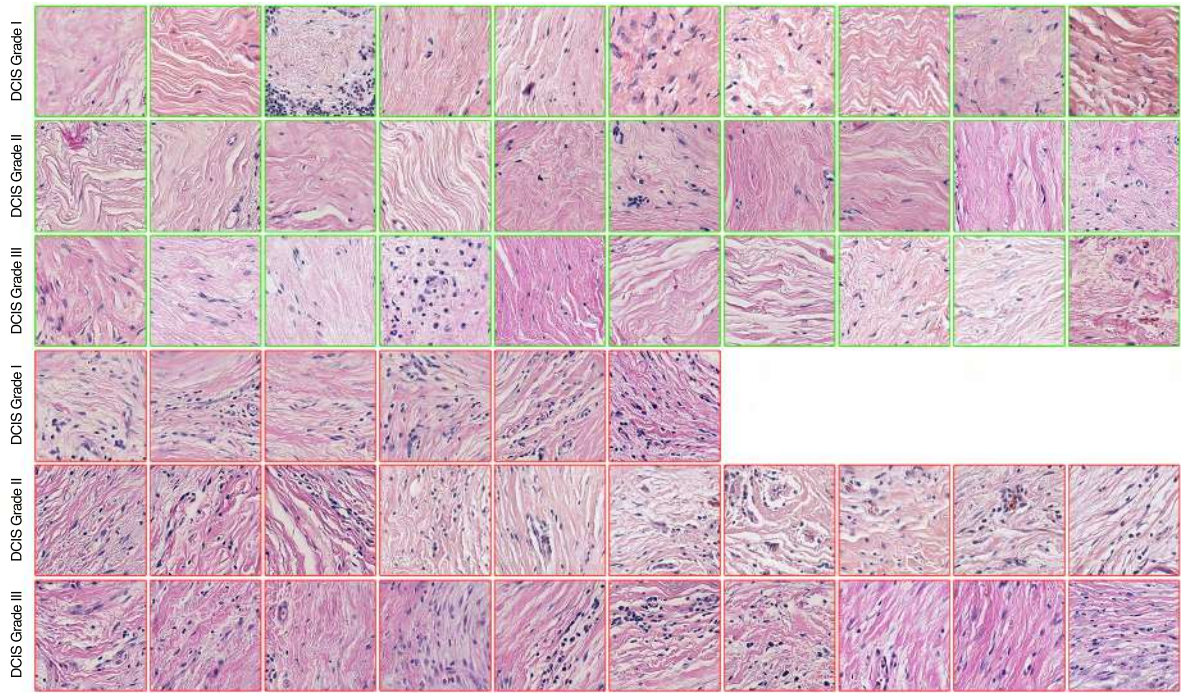

Figure 8.5: Examples of stroma patches for different grades of DCIS with low and high probabilities for tumor-associated stroma. The first three rows show image patches (shown in green bounding boxes) for different grades of DCIS that have low probabilities for tumor-associated stroma. The image patches in rows four to six (shown in red bounding boxes) show example patches for different grades of DCIS with high probabilities for tumor-associated stroma. Note that each patch is taken from an independent DCIS lesion; however, due to the lack of DCIS grade 1 lesions with high tumor-stroma probabilities, the examples in the fourth row are taken from two lesions only. Subjectively, cellularity generally appears higher in grade 2 and 3 lesions.

criteria for classifying stromal changes have not been developed and used clinically. Accordingly, agnostic approaches, such as using deep learning techniques, are well-suited to investigating the morphology of breast stroma because visual characterization or feature selection is not required.

In the surgical management of breast cancer, it may be important to excise malignant epithelium and tumor-associated stroma. The ability of the system to objectively identify regions of altered stroma associated with tumor may additionally complement the pathologist's diagnosis and may assist in identifying stromal tissue that should be included in tumor margins.

Using our proposed system, we achieved an AUC of 0.962 for breast cancer diagnosis. The results demonstrate that breast cancer can be accurately diagnosed based on the analysis of stromal features alone, suggesting the centrality of alterations to

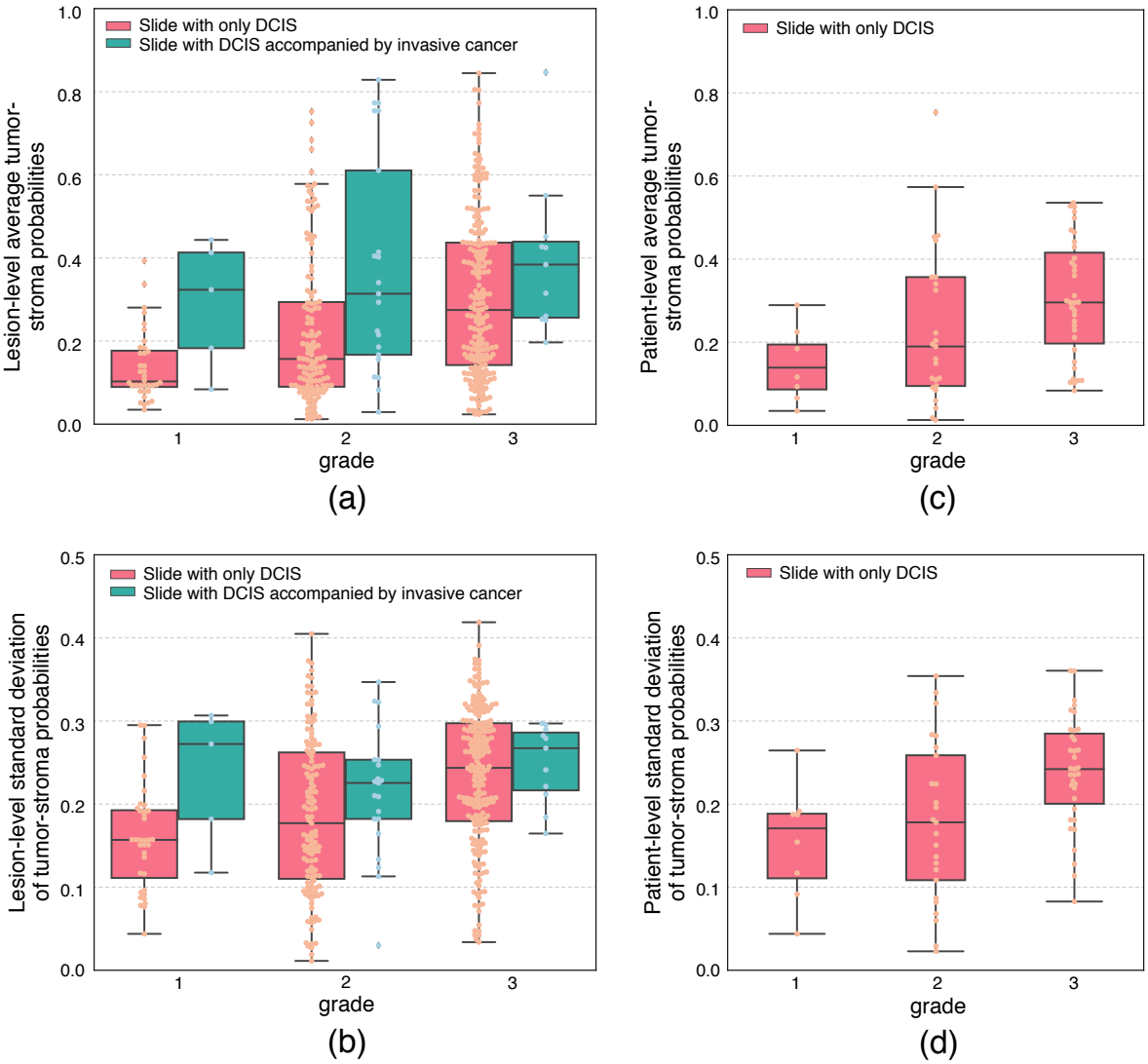

Figure 8.6: Boxplots for the mean and standard deviation of tumor-associated stroma probabilities for the pixels at a distance of $< 175\mu m$ from the DCIS margin for different histological grades of DCIS. (a) and (b) show the mean and standard deviation of tumor-associated stroma probabilities at the lesion level. The orange boxplots show the results for all DCIS foci in slides diagnosed as DCIS (n=375). The green boxplots show the results for slides containing DCIS accompanied by invasive cancer (n=37). (c) and (d) show the mean and standard deviation of tumor-associated stroma probabilities at the patient level (n=58) for slides with DCIS diagnosis.

the breast stroma in the process of breast carcinogenesis. The ability of the system to objectively identify regions of stroma affected by the tumor can additionally assist the pathologist in determining tumor margins and its associated zones of influence in breast tissue resections.

A key goal of our project was to use an unbiased data-driven approach to examine potential relationships between the patterns of stroma surrounding DCIS lesions

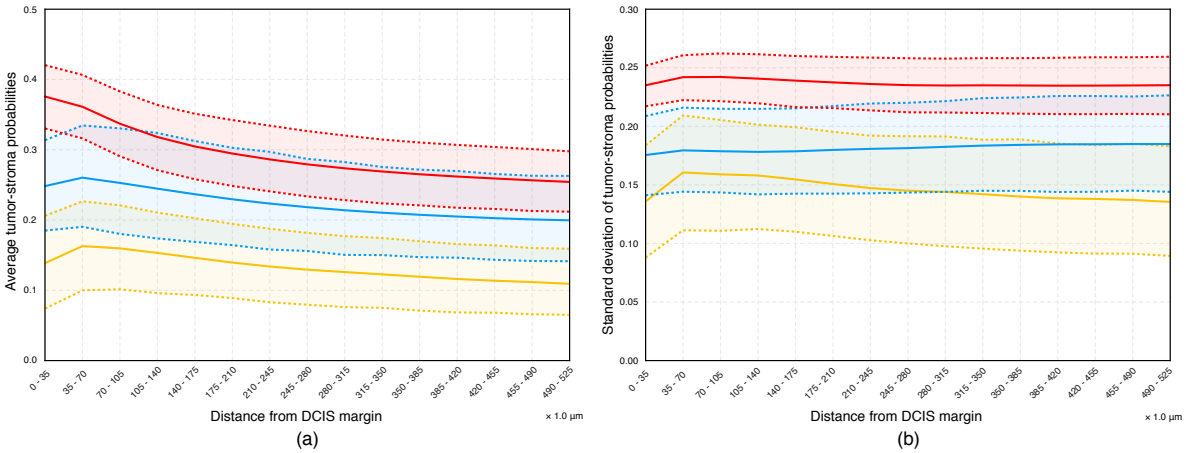

Figure 8.7: Plots of the mean (a) and standard deviation (b) of tumor-associated stroma probabilities for different ranges of distances from the DCIS margin for different histological grades. For each measure (mean and standard deviation of probabilities), the curve for each grade is constructed using the mean along with the 95% confidence intervals around the mean of the population.

and DCIS grade. It is hypothesized that transformation of the stroma starts in an early phase of DCIS development[181–183], and there is growing evidence that stroma contributes importantly to the transformation of DCIS to invasion[181,182,185]. Thus, we tested the hypothesis that stromal alterations may serve as a proxy for the potential for DCIS to undergo an invasive transformation.

Although we did not train our model on DCIS, we found that tumor-associated stroma probabilities were significantly higher in grade 3 DCIS, with the amount of tumor-associated stroma generally increasing with increasing lesion grade. Although we were unable to distinguish pathologically defined DCIS grade 2 from DCIS grade 1 or grade 3, data show that reproducibility of DCIS grade 2 is poor[190], suggesting that this comparison may have limited value. Despite this limitation, data show that high-grade DCIS may have a higher risk of recurrence after surgical excision than low-grade DCIS, and when recurrences occur after DCIS treatment[195–197], they occur earlier for higher grade lesions[197–199]. Studies also suggest that occult invasion is more common among women with image-guided biopsies diagnosed with higher grades of DCIS[196,200] and that this may be important because grade of invasive cancer is generally matched with grade of accompanying DCIS[198,201]. Further, a low percentage of high-grade DCIS has been associated with positive axillary nodes or later metastases, suggesting that at least a subset of such lesions are associated with occult invasion or disseminate through an undefined mechanism. Finally, ongoing prospective trials (LORIS[202], LORD[203], and COMET[204]) are assessing conservative management of low-risk DCIS, given indirect evidence that

many such lesions will never cause harm during a woman's lifetime. Our data suggest that consideration of evaluating stromal changes to assess its role as a potential biomarker of risk for recurrence may have value in such trials.

There are several limitations to our study. Our dataset was limited to one study population, thus repeating this analysis in other populations is important. Additionally, this study was limited to breast tissue sections obtained at time of biopsy and further insights might be obtained by assessing the stromal patterns on WSIs from subsequent matched breast tissue surgical resections. Although our comparison of tumor-associated stroma in pure DCIS and DCIS with invasive cancer attempted to focus on areas of slides that were further away from invasive cancer, because of limited amount of tissue in some biopsies, there was a risk that stromal changes associated with some DCIS areas reflected nearby invasion. Analyzing WSIs of resected specimens would help alleviate this risk, provided that avoiding changes associated with the prior biopsy site does not pose insurmountable challenges. Further experiments on larger cohorts of DCIS with long-term clinical follow-up are needed, to study the potential that stromal features may have prognostic value. For example, stromal analysis may help define which DCIS, grade 2, will behave indolently like grade 1 versus more aggressively like grade 3.

In conclusion, we have developed a deep learning approach utilizing CNN to identify the presence of cancer in WSIs based on tumor-associated stromal alterations in diagnostic image-guided breast biopsies. Further, we demonstrated that deep learning techniques can define stromal features that are related to DCIS grade. Additional studies using these approaches with follow-up of DCIS cases may be useful.

Acknowledgment

This project was funded in part by the Intramural Research Program of the National Cancer Institute, National Institutes of Health, Bethesda, Maryland and a competitive award to MES and LAB funded through the sale of breast cancer awareness postage stamps. The authors wish to acknowledge the financial support by the European Union FP7 funded VPHPRISM project under the grant agreement n°601040. Pamela Vacek and Donald Weaver are currently funded under a U01 exploring stromal contributions to tumor progression (U01 CA196383).

Supplementary material

Convolutional neural network (CNN) architecture

CNN I and *CNN II* are based on the VGG architecture[135]. Details of our network configuration are presented in Table 8.2. VGG-net uses 3×3 filters throughout the convolutional layers of the network. Although 3×3 convolutions have very small receptive fields, a stack of multiple 3×3 convolutions without any max-pooling in between can effectively yield a larger receptive field (e.g. three such layers have a 7×7 effective receptive field). The real advantage lies in the fact that having multiple non-linearities after each layer in a stack, can offer a model with more discriminative power.

A naive approach to classifying the entire pixels in the gigapixel WSI is to apply the trained CNN to every single pixel in the WSI in a sliding window fashion. This approach is highly inefficient and slow at inference because input patches to the CNN have huge overlap, leading to a significant amount of redundant computations. To overcome this, we replaced the two fully connected layers by convolution operations with kernels of size 1×1 spatial extent[205]. Subsequently, this enabled us to apply our network to much larger images at test time, as convolution operations can be applied to the full extent of the image, producing a dense prediction map in the output layer, with one spatial location for each field of view of input. However, the use of pooling layers and convolutions in valid mode (in which convolution is only computed where the input and the filter fully overlap) results in output prediction maps that are at a lower resolution than the input. We simply used the nearest neighbor resampling method to upsample the output map to match the size of the input image.

Preprocessing of WSIs and ground truth ROIs

The tissue segmentation algorithm previously described[206] was used to exclude background (area on slide where there was no tissue) from our analyses. We generated WSI labeled masks from the annotations. RGB patches were subsequently randomly generated from the original WSI in masked areas, on the fly, during training and validation. Preprocessing the extracted patches involved scaling between 0 and 1, and subtracting the mean for each channel, calculated on the training set.

Table 8.2: The architecture of the networks used in this study. Convolutional neural network I *CNN I* was used for classification of breast tissue into epithelium, stroma, and fat. *CNN I* was used for identification of tumor-stroma regions. *CNN III* was used for WSI classification of breast biopsies into normal/benign and invasive cancer. The weights of the blocks in grey were taken from the previously trained *CNN II* network. Inside the table Conv3@N is a convolutional layer with a kernel size of 3×3, and N filters. Batch normalization and ReLU precede each convolution.

CNN I	CNN II	CNN III
input (224×224)		8 input patches (336×336)
conv3@12 conv3@12	conv3@64 conv3@64	conv3@64 conv3@64
max pooling		
conv3@24 conv3@24	conv3@128 conv3@128	conv3@128 conv3@128
max pooling		
conv3@48 conv3@48	conv3@256 conv3@256 conv3@256	conv3@256 conv3@256 conv3@256
max pooling		
conv3@128 conv3@128 conv3@128	conv3@512 conv3@512 conv3@512	conv3@512 conv3@512 conv3@512
max pooling		
-	conv3@512 conv3@512 conv3@512	conv3@512 conv3@512 conv3@512
-	max pooling	
FC-2048	FC-4096	conv3@256, /2
FC-2048	FC-4096	global max pooling
		concatenation layer
		FC-2048
		FC-1024
soft-max		

Training procedure

All the CNNs in our work were trained using stochastic gradient descent with Nesterov momentum[207]. The mini-batch size was set to 150 and 22 for *CNN I* and *II*, respectively, and 6 for *CNN III*. An adaptive learning rate scheme was utilized for all networks with an initial learning rate of 0.01 for *CNN I* and *II* and 0.005 for *CNN III*. The learning rate was decreased by a factor of 5 if no increase in the performance was observed on the evaluation set, over a predefined number of epochs (epoch pa-

tience), which was initially set to 10 and increased by 20% following each drop in the learning rate.

The weights of our networks were initialized using the strategy in[137]. For all networks, we used batch normalization[208] immediately after each convolution layer, followed by ReLU[209] as nonlinearity. Weight decay was set to 0.003 and 0.0001 for CNN *I* and *II*, respectively, and 0.0005 for *CNN III*, along with 50% dropout[210] on convolutional layers replacing fully connected layers in all networks.

To maximize the amount of information extracted from the dataset and make the developed CNNs more robust to common variations, the dataset was augmented by random flipping and rotations as well as by random jittering of the hue and saturation of pixels in the HSV color space.

Training of *CNN I* and *II* involved two rounds of hard sample mining[211]. Initial training of *CNN I* was based on a limited set of annotations made on a subset of the training data. New training data were generated after each hard sample mining round but annotating the misclassified areas of WSIs in the training set. For *CNN II*, this step was equivalent to hard-negative mining. False positive stromal regions detected in the normal WSIs within the training set were automatically added to the training set following each hard-negative mining round.

We used a weighted cross-entropy loss for the training of *CNN II*. The reason for this is that the amount of annotated tumor-associated stroma was significantly less than the normal stromal available in the training set. On the other hand, because of the wide spectrum of benign abnormalities of breast tissue, normal stroma may reveal a more heterogeneous appearance. To ameliorate this problem, we generated balanced mini-batches to increase the network's capacity in learning discriminative features for the minority class. However, because the class distribution in each mini-batch did not represent the actual skewed distribution of the data, we gradually increased the weight for misclassification of the samples of the normal/benign class in the loss calculation. This is achieved by applying an exponential growth function to a provided loss weight. We set the initial loss weight for the normal/benign samples to 1 and used a growth rate of 1.0034.

Region selection for CNN III

To predict the label of the WSI image, we trained a CNN to directly predict the WSI label (diagnosis) in its output (Fig. 8.1). The system obtains the dense prediction results produced by *CNN II*. Next, 8 sub-regions of size 336×336 were selected from the WSI, representing regions with high likelihoods for cancer-associated stroma. These patches were conditioned to be composed of at least 95% stromal pixels. The next

section describes the procedure for selecting these patches. The feature maps produced by the antepenultimate layer of *CNN II* for each of these layers were then fed into a small CNN, which we denote as *CNN III*. Details of this network are presented in Table 8.2. The feature vectors of size 1×256, generated after applying global max pooling on the output of the last convolutional layer of this network, were concatenated for the 8 input patches and subsequently passed to two fully connected layers before being fed into a softmax classifier for predicting the WSI score.

Selection of candidate stromal regions as input to *CNN III*

Candidate stromal regions were identified by applying iterative thresholding on the probability map produced by *CNN II* to find 16 connected components. The initial threshold was set to 0.9 to generate tumor-stroma connected components. Connected components were sorted based on their size and the 16 largest connected components were selected as candidate regions. The threshold was iteratively reduced by 0.1 until the condition of finding 16 connected components was met. Eight patches were then randomly selected from these regions and fed to *CNN III*. In case there exist no connected components for a WSI (e.g. normal slides with no suspicious stromal regions), 8 zero-valued images were fed as input to *CNN III*.

Note that to generate the final score for the WSI, we took the average probabilities produced by *CNN III* for two sets of 8 randomly selected patches from these connected components.

Summary and conclusions

9

9.1 Thesis Summary

In **Chapter 1** we presented background material on the application of machine learning techniques for analysis of image data. We identified breast cancer as a leading cause of cancer death among women worldwide and discussed the potential benefits of computerized diagnostic systems based on machine learning for improving histopathological diagnostics. Deep learning, a state-of-the-art form of machine learning, was introduced and its use was motivated for application to analysis of breast histopathology images. The primary aim of the thesis was stated as:

> "to develop automated systems for analysis of H&E stained breast histopathological images. This involved automatic detection of ductal carcinoma in-situ (DCIS), invasive, and metastatic breast cancer in whole-slide histopathological images."

The secondary aim of the thesis was stated as:

> "to identify new diagnostic bio-markers for the detection of invasive breast cancer in H&E stained breast histopathological images."

In **Chapter 2** we presented data on the sources of variation of the widely used H&E histological staining. The chapter also presented a new algorithm to reduce these variations in digitally scanned tissue sections.

In **Chapter 3** we presented a fully automated algorithm for standardization of whole-slide histopathological images to reduce the effect of staining variations. To the best of our knowledge this is the first algorithm for standardizing whole-slide H&E stained histopathological images. The proposed algorithm, called whole-slide image color standardizer (WSICS), utilized color and spatial information to classify the image pixels into different stain components. The chromatic and density distributions for each of the stain components in the hue-saturation-density color model were aligned to match the corresponding distributions from a template whole-slide image (WSI) to yield the standardized image.

In **Chapter 4** a new algorithm for automatic detection of regions of interest in whole-slide histopathological images was proposed. The algorithm generated and classified superpixels at multiple resolutions to detect regions of interest. The algorithm emulated the way the pathologists examine the whole-slide

histopathology image by processing the image at low magnifications and performing more sophisticated analysis only on areas requiring more detailed information.

In **Chapter 5** we presented and evaluated a fully automatic method for detection of ductal carcinoma in situ (DCIS) in digitized H&E stained histopathological slides of breast tissue. The proposed method used the algorithm described in Chapter 4 to detect epithelial regions in whole-slide images (WSIs). Subsequently, spatial clustering was utilized to delineate regions representing meaningful structures within the tissue such as ducts and lobules. A region-based classifier employing a large set of features including statistical and structural texture features and architectural features was then trained to discriminate between DCIS and benign/normal structures.

In **Chapter 6** we presented context-aware stacked convolutional neural networks (CNN) for classification of breast WSIs into normal/benign, DCIS, and invasive ductal carcinoma (IDC). We first trained a CNN using high pixel resolution to capture cellular level information. The feature responses generated by this model were then fed as input to a second CNN, stacked on top of the first. Training of this stacked architecture with large input patches enabled learning of fine-grained (cellular) details and global tissue structures.

In **Chapter 7** we presented results of the CAncer MEtastases in LYmph nOdes challeNge (CAMELYON16), established to objectively compare the performance of automated algorithms for detecting metastases in whole slide images. The performance of the algorithms was tested against a panel of 11 pathologists in a simulated diagnostic setting and against a single pathologist in a setting without time constraint. The results showed that some of the deep learning based systems achieved performances on par with a pathologist scoring without time constraint and significantly outperformed all the pathologists in the simulated diagnostic setting.

In **Chapter 8** The interplay between malignancies and their microenvironment is now recognized as a pivotal factor in tumor growth and metastatic potential. The microenvironment around the tumor, in particular, stromal cells, undergo drastic changes during tumor growth. In this study, we applied machine learning techniques to automatically assess these alterations in scanned tissue sections. In the first stage, we developed an algorithm based on deep convolutional neural networks for classification of breast specimen biopsies purely based on stromal characteristics of the breast tissue. Next, the proposed system

was applied to DCIS lesions to assess the potential of such lesions to establish a microenvironment which supports invasive progression.

9.2 Key contributions and findings

The techniques presented in this thesis represent the evolution of my research over the last 4 years. The ordering of the chapters is chronologically in line with advances in our machine learning methodologies for analyzing WSIs of breast tissue, starting from usage of conventional machine learning in Chapters 4 and 5 to employing state-of-the-art deep learning techniques in Chapters 6−8.

Studying the sources of staining variations and development of the first WSI standardization algorithm

Our journey began by studying the major sources of staining variations in a large study where we collected data from 5 laboratories in the Netherlands. We showed that staining protocols in different laboratories and staining on different days of the week were the major factors causing color variations in histopathological images (Chapter 2). While pathologists can effectively cope with such variations, the performance of CAD systems can be hampered by them. In this chapter we also designed an algorithm to reduce the effect of these variations by standardizing small image patches. There were, however, two major limitations for this algorithm. Firstly, it could not be used for standardizing giga-pixel WSIs. Secondly, its performance could deteriorate when there was a significant imbalance in the amount of eosin and hematoxylin staining in the image.

The requirement to develop high throughput algorithms to process giga-pixel histopathological WSIs provided the motivation to further investigate a new approach for standardization which could operate on WSIs and cope with all possible sources of staining variations. As such, in Chapter 3 we introduced the first algorithm for standardizing whole-slide H&E stained histopathological images, called WSICS. One major common step among the algorithms for standardization of histopathological images is identification of the stain components in the image prior to standardization. Unlike previously published algorithms which rely solely on color information to identify stain components WSICS incorporated spatial information which made it significantly more robust. In our evaluations, we showed that the performance of pre-existing CAD systems could be significantly improved by utilizing our algorithm. While the algorithm was designed in the era that conventional machine learning was the most popular choice for development of computerized diag-

nosis systems, we also found that the performance of convolutional neural networks could be significantly improved by its use. Ciompi et al.[145] showed that the performance of a CNN for classification of colorectal tissue on an independent dataset with different staining appearance could be improved from an accuracy of 51% to roughly 80% following the utilization of WSICS algorithm. WSICS was also used by Wang et al.[34], the winner of the CAMELYON16 challenge and was a major contributing factor to improve the ranking of their system from the 4th position to the 1st.

Development of a DCIS detection system

DCIS is a clinically challenging subcategory of breast proliferative lesions that is noninvasive and is considered a precursor for invasive ductal carcinoma (IDC). It encompasses a heterogeneous group of lesions with highly variable morphology, biomarker expression, genomic profile, and natural progression[85]. Development of a DCIS detection system for analysis of WSIs was challenging for two major reasons: (1) WSIs are large and may contain hundreds of structures which need to be analyzed. Therefore, obtaining a small false positive rate while still retaining a high sensitivity can be hard, and (2) A CAD system that operates on the WSI level should be able to handle a larger set of heterogeneous benign structures (e.g. adenosis, UDH, cysts, etc.) and artifacts (due to staining/cutting) to detect DCIS.

As the first step of our DCIS detection system, we introduced a multi-scale superpixel classification approach for finding epithelial areas in the WSI. This algorithm enabled subdivision of the WSI into regions which adapt to the underlying image data, such that every superpixel was mostly homogeneous, thereby facilitating the classification of the tissue components within the WSI. Detection and contouring of diagnostically relevant regions was based on a spatial clustering approach operating on the graphs built on the centroids of epithelium labeled superpixels. A set of texture-based and architectural features were then extracted from the delineated epithelial structures for subsequent classification into normal/benign lesions and DCIS using a gradient boosted classifier.

Evaluation was conducted both on the slide and the lesion level using FROC analysis. The results show that to detect at least one true positive in every DCIS containing slide, the system finds 2.6 false positives per WSI. The results of the per-lesion evaluation show that it is possible to detect 80% and 83% of the DCIS lesions in an abnormal slide, at an average of 2.0 and 3.0 false positives per WSI, respectively. To the best of the authors' knowledge, this is the first fully automated DCIS CAD system that operates at the WSI level and has been evaluated on a dataset collected from routine clinical practice.

Although our system was evaluated on an independent dataset collected from

routine clinical practice, further evaluations are necessary on additional independent cohorts of breast cancer patients that have different staining characteristics to those of our training dataset.

We believe the biggest impediment to our method is its computational expense. A C++ implementation of our system running on a laptop with an Intel Core i7 CPU (2.4 GHZ) and 16 GB of Ram takes between 30 to 55 minutes for processing a WSI.

Context-aware stacked convolutional neural networks

The primary aim of my thesis was to develop automated systems for classification of breast histopathology images into normal/benign, DCIS and invasive carcinoma. Following the development of our DCIS detection system, a logical step was to evolve the system to allow for detection of invasive cancer. Recent advances in neural network based artificial intelligence, popularly known as deep learning, motivated us to solve this task with deep convolutional neural networks.

The main challenge in the design of our classification framework was that the appearance of many benign diseases of the breast (e.g. usual ductal hyperplasia) mimic that of DCIS, hence requiring accurate texture analysis at the cellular level. Such analysis, however, is not sufficient for discrimination of DCIS from IDC. DCIS and IDC may appear identical on cellular examination but are different in their growth patterns which can only be captured through the inclusion of larger image patches containing more information about the global tissue architecture. Because of computational constraints, however, it is not feasible to train a deep CNN with large patches at high resolution that contain enough context.

To tackle this we introduced context-aware stacked convolutional neural networks. For the development of this model we first used a deep CNN employing high pixel resolution information to classify the tissue into different classes. To incorporate more context to the classification framework, we fed a much larger patch to this model at test time. The feature responses generated by this model were then input to a second CNN, stacked on top of the first. This stacked network uses the compact, highly informative representations provided by the first model, which, together with the information from surrounding context, enables it to learn the global interdependence of various structures in different lesion categories.

Comparing the results of our network to a conventional CNN architecture showed that the performance is significantly improved. Our system achieved an AUC of 0.962 for the binary classification of normal/benign slides from cancerous slides and an accuracy of 81.2% for the 3-class classification problem.

Although our primary aim was to facilitate pathology diagnostics by discriminating between different breast lesion categories, our system could serve as an im-

portant first step for the development of systems that aim at finding prognostic and predictive biomarkers within malignant lesions[146].

CAMELYON16: Challenge on detection of breast cancer metastases in sentinel lymph nodes

The lymph node status is an important prognostic factor for breast cancer patients, used for therapy planning and patient management. Assessment of the lymph nodes by pathologists is a laborious and, to some extent, subjective task. Availability of an automated system to support such diagnostics will be beneficial to patients because of improved accuracy. Additionally, computerized diagnostics will improve efficiency, reduce costs and will alleviate the lack of experienced pathologists.

To further investigate this, and find what the current state of the art is, we organized the CAMELYON16 grand challenge[129] from November 2015 to November 2016 in conjunction with and with the support of the 2016 IEEE International Symposium on Biomedical Imaging. The aim of our challenge was to assess the performance of automated machine learning systems in detecting metastases in H&E stained tissue sections of lymph nodes of breast cancer patients and compare it to pathologists in a diagnostic setting.

The challenge was highly successful, with over 30 submissions from all around the world. The results showed that deep learning operates at the level of a pathologist without time constraint and that it achieves significantly better performance than all the pathologists in the diagnostic setting. This study shows, for the first time, that deep learning based methods are able to solve clinically relevant tasks in diagnostic pathology at the level required for routine diagnostics.

We believe this study will accelerate progress in the field of computerized pathology diagnostics. CAMELYON16 stimulated many research groups and companies to, for the first time, focus on application of machine learning to histopathology. Our challenge attracted lots of media attention, with articles being published on over 30 websites and the challenge being mentioned in the White House report on "The national AI research and development strategic plan 2016".

Identification and characterization of tumor-associated stromal alterations using deep neural networks

The secondary aim of this thesis was to identify new diagnostic bio-markers for the detection of invasive breast cancer in H&E stained breast histopathological images. We focused on the analysis of tumor induced stromal alterations which is recognized as a critical determinant of tumor progression and therapeutic response. It has been

shown that for breast ductal carcinoma, gene expression in the stroma adjacent to tumor strongly correlates with prognosis. Tumor-associated stromal alterations have a morphological substrate, referred to by pathologists as desmoplastic alterations. Pathologists are, however, incapable of accurately identifying and characterizing tumor induced stromal alterations. In this study we aimed to apply deep learning techniques to automatically assess these alterations in scanned tissue sections.

Development of robust computerized algorithms for discriminating patterns of normal stroma and tumor-associated stroma in histopathology images is a complex task mainly because there is no precise definition of the visual characteristics of tumor-associated stroma. Machine learning algorithms could be leveraged for the classification task. However, conventional algorithms require careful engineering of domain specific visual features specified by experts. Deep learning techniques, in contrast, obviate the need for feature engineering by learning the most predictive features directly from the images, given a large data set of labeled examples.

Our proposed approach for classification of the WSI into benign/normal or invasive cancer had multiple stages. At first, a CNN was trained to classify the WSI into 3 major tissue components: fat, stroma, and epithelium. Subsequently, we limited all further analyses to stromal regions of the tissue only. Next, using a cascade of two deep learning models, a system was proposed to classify an entire WSI into normal/benign or invasive breast cancer. This system could be deployed to directly predict the WSI score purely based on stromal characteristics of the tissue. The first part of the model was trained to discriminate normal and tumor-associated stroma at the patch level. The second stacked network received the feature maps generated by the first network for 8 hotspot regions with high probabilities for tumor induced stroma. This stacked architecture directly predicts the label for the WSI.

We validated our network by showing its ability to identify cancerous tissue sections exclusively on the basis of stromal appearance, without having access to epithelial/tumor features. The proposed model achieved an AUC of 0.962 for breast cancer diagnosis. The system was subsequently applied to DCIS lesions, to assess the potential of such lesions to establish a microenvironment which supports invasive progression. We observed an increasing amount of tumor associated stroma with increasing grades of DCIS, which also extended to a larger distance from the lesions. DCIS lesions in slides that also contained an invasive component generally possessed higher amounts of tumor associated stroma compared to DCIS-only slides.

Our findings suggest that our proposed system may be used as a strong diagnostic and potentially prognostic tool for breast cancer in clinical practice. Firstly, the proposed system may be used as an objective, quantitative tool to aid the patholo-

gists in histological grading of DCIS. The use of our system can provide complementary information to the pathologist for more objective assessment of the lesion. Our system can also be combined with computerized DCIS detection systems[86,87,114] to yield a more accurate automated grading of DCIS. Secondly, studies to understand the molecular and genetic gene expression patterns and phenotypes of breast cancer microenvironment could benefit from our tool by demarcating the stromal areas associated with the tumor. Finally, the proposed stroma assessment tool may prove useful for extraction of prognostically important information from tumor biopsies.

9.3 Opportunities for further research

The field of computerized analysis of breast histopathology images for breast cancer diagnosis is an exciting one. While considerable amount of work has been done towards the aims of this thesis, there is still the need for further research to answer several fundamental questions.

Role of stain standardization in application of machine learning algorithms to histopathology

It was shown that our proposed standardization algorithm improves the robustness of both conventional machine learning algorithms[96] and deep convolutional neural networks[34,145]. Both Ciompi et al.[145] and Wang et al.[34] achieved their best results after standardizing both the training and test sets. The requirement to standardize test set images, however, comes at the cost of increased computation times. Therefore, one important pathway for future work is to develop machine learning pipelines which is implicitly invariant to the staining variations of the data. Of particular interest is the possibility of incorporating the domain adaptation theory[212,213] into the training framework of the deep learning model. As such, the focus in training is on learning features that combine discriminativeness with domain invariance. Ganin et al.[214] proposed a general framework for training deep models that enable jointly optimizing the underlying features as well as two discriminative classifiers operating on these features: (i) the classifier for predicting the label of the input image, and (ii) the domain classifier that discriminates between the source and target domains during training. Within this framework, our standardization algorithm could be used to generate a rich set of adversarial examples to the source domain. This could be achieved by applying transformations using the look-up-tables learned for matching different domain, on the fly, during training.

Incorporating high-resolution context for the classification task

In Chapter 6 we presented context-aware stacked convolutional neural networks for the classification of breast specimen into normal/benign, DCIS, and IDC. Our approach is of particular interest when there is a requirement for high-resolution information as well as large context for the classification task. An alternative avenue of research is the use of U-Net architecture[142] which allows the seamless segmentation of arbitrary large images. However, to obtain a receptive field of approximately 768×768 pixels which was obtained by our model, we need a U-Net architecture with 27 layers (assuming a network with repeated application of two 3×3 convolutions and a 2×2 max pooling with a stride of 2 according to the original U-Net design). Training this network may therefore be more memory-demanding than our approach. Further research is needed to investigate this.

Prognostic and predictive value of tumor-associated stroma

Further research is needed to determine the value of tumor-associated stroma to build prognostic and predictive models for breast cancer patients. Of particular interest is the possibility for developing a system for the assessment of the risk for invasive cancer development among patients diagnosed with DCIS or breast benign disease (BBD). Both of these research avenues require large datasets of breast cancer patients with follow-up data.

In chapter 8, we followed a supervised learning approach using CNN for analyzing stromal patterns of breast tissue. Training samples for the positive class were obtained by annotating examples of peri-tumor stromal regions in slides containing invasive cancer. One direction will be to use the developed CNN for the analysis of the patterns of stroma among patients diagnosed with DCIS or BBD to predict the risk of future invasive cancer development among them. This approach, however, is based on the assumption that similar stromal patterns to invasive tumor-stroma may appear in patients with DCIS or BBD who are at higher risks for development of invasive cancer. This may not necessarily hold true.

An alternative approach is to endow computers with an understanding of the stromal patterns in breast tissue specimen regardless of their potential association to a higher or lower risk for invasive cancer development (unsupervised feature learning). Generative models are currently one of the most promising approaches towards this goal. Generative Adversarial Networks (GANs)[215] have recently emerged as a powerful framework for learning generative models that map samples from a simple latent distribution to a particular data distribution of interest. Intuitively, a model cable of generating arbitrary data distributions may serve useful feature representa-

tions for auxiliary problems. As an example, a GAN trained to generate convincing examples of various patterns of stroma may have learned the feature representations that are discriminative for predicting the risk for future development of invasive cancer. Despite producing impressive results[216,217], in their current form, GANs offer no straight forward path for projecting back the data into latent space which could be used as feature representations. Recently, Donahue et al.[218] proposed Bidirectional Generative Adversarial Networks (BiGANs) as a means of learning the inverse mapping of the real data distribution into the latent feature space of the generative model, and demonstrated that the resulting learned feature representation is useful for auxiliary supervised discrimination tasks.

One interesting future work avenue will be the use of an unsupervised feature learning framework such as BiGANs for building a model to predict the risk for invasive cancer development among patients diagnosed with DCIS or BBD.

The influence of large high-quality data sets for conducting influential studies

The continuing generation of high-quality standardized datasets from large populations of cancer patients will be critically important to compare, interpret or validate experimental data and computerized tools. Grand challenges providing large datasets and benchmarking methods are powerful means to compare varying approaches in a fair and reproducible way. A grand challenge will ensure that the largest possible community of researchers is able to address the most important problems in cancer medicine today. We believe CAMELYON16 challenge accelerated progress in the field and stimulated lots of researches from all over the world to work on giga-pixel pathology images for cancer diagnostics. Future organization of such challenges with standardized high quality datasets could instigate a much wider proliferation of machine learning solutions to problems in pathological diagnostics.

Samenvatting

Hoofdstuk 1 geeft achtergrondinformatie over de toepassing van kunstmatige intelligentie bij de analyse van digitale beelden. We hebben borstkanker gedentificeerd als n van de belangrijkste kanker-gerelateerde doodsoorzaken van vrouwen wereldwijd. Dit hoofdstuk beschrijft de mogelijkheden om de histopathologische diagnostiek te ondersteunen door het gebruik van kunstmatige intelligentie. Tevens wordt een relatief nieuwe vorm van kunstmatige intelligentie, gebaseerd op diepe artificile neurale netwerken ('deep learning'), gentroduceerd en wordt de bruikbaarheid voor analyse van borstkankerhistopathologie uitgelegd. Er wordt hierbij gebruik gemaakt van volledig gedigitaliseerde weefselcoupes ('whole slide images'; WSI).

De primaire doelstelling van dit promotieonderzoek was:

> "De ontwikkeling van algoritmen voor de volledig automatische analyse van hematoxyline-eosine (H&E) gekleurde borstweefselcoupes. Dit omvat de automatische detectie van ductale carcinoma in-situ (DCIS), invasieve en metastatische kanker in WSI."

De secundaire doelstelling van dit promotieonderzoek is:

> "Het identificeren van nieuwe diagnostische biomarkers voor de detectie van invasieve borstkanker in H&E-gekleurde, histopathologische weefselcoupes."

In **Hoofdstuk 2** worden data gepresenteerd die de verschillende variatiebronnen beschrijven van de meeste gebruikte weefselkleuring in de pathologie: de H&E kleuring. Dit hoofdstuk beschrijft ook een nieuw algoritme om deze variaties in WSI te reduceren.

In **Hoofdstuk 3** wordt een algoritme beschreven dat volledig automatisch WSI kan standaardiseren om problemen ten gevolge van kleuringsvariaties te verminderen. Zover ons bekend is dit het eerste algoritme dat geheel automatisch een WSI van een H&E-gekleurde weefselcoupe kan standaardiseren. Het algoritme, 'whole-slide image color standardizer' (WSICS) genoemd, gebruikt zowel kleur- als spatile informatie voor het herkennen van pixels die uit de verschillende kleuringscomponenten bestaan. De kleur- en dichtheidsverdelingen voor elk van deze kleuringscomponenten in het hue-saturation-density kleurmodel worden vervolgens dusdanig getransformeerd dat ze de gewenste standaardverdelingen benaderen en daarmee het beeld standaardiseren.

In **Hoofdstuk 4** wordt een nieuw algoritme voor de automatische detectie van potentieel interessante gebieden in WSI beschreven. Het algoritme genereert en classificeert zogenaamde superpixels op verschillende resoluties. Dit algoritme bootst de manier na waarop een patholoog een weefselcoupe onder een microscoop bekijkt,

door het beeld eerst op lage vergroting te analyseren en vervolgens meer verfijnde analyse te doen van alleen die gebieden waarvoor dat nodig is.

In **Hoofdstuk 5** presenteren en evalueren we een methode voor de volledig automatische detectie van DCIS in WSI van borstweefsel. De voorgestelde methode gebruikt het algoritme dat beschreven wordt in Hoofdstuk 4 om gebieden met epitheel te herkennen. Vervolgens wordt er een spatile clustering van de superpixels uitgevoerd om automatisch relevante weefselstructuren te omlijnen, bijvoorbeeld ducti en lobuli (melkgangen en -klieren). Een statistisch beslismodel wordt ontwikkeld die op basis van een groot aantal kwantitatieve eigenschappen (statistische en structurele textuur eigenschappen en architecturale eigenschappen) van de herkende structuren het onderscheid kan maken tussen gebieden met DCIS en benigne structuren.

In **Hoofdstuk 6** presenteren we context-gevoelige, gekoppelde convolutionele neurale netwerken (CNN) voor de classificatie van WSI van borstcoupes in benigne laesies, DCIS en invasieve tumor. Allereerst hebben we een CNN getraind die op basis van hoge-resolutie beeldinformatie een analyse doet van eigenschappen op celniveau. De resultaten van deze CNN worden vervolgens gebruikt als invoer voor een tweede CNN, die gekoppeld is aan de eerste CNN. Het direct trainen van deze gekoppelde architectuur met grotere deelgebieden van een WSI maakt het mogelijk om tegelijkertijd zowel details op celniveau als globale architecturale informatie te gebruiken.

In **Hoofdstuk 7** presenteren we resultaten van de CAncer MEtastases in LYmph nOdes challeNge (CAMELYON16) competitie. Dit is een competitie die wij hebben georganiseerd met als doel het op objectieve wijze vergelijken van de prestaties van algoritmes die automatisch lymfkliermetastasen in WSI van borstkanker patinten kunnen detecteren. De resultaten van de algoritmen zijn vergeleken met die van een panel van 11 pathologen in een gesimuleerde diagnostische setting en met die van n patholoog die zonder enige tijdsbeperking de coupes beoordeeld heeft. De resultaten van onze studie laten zien dat een aantal van de op 'deep learning'-gebaseerde algoritmen even goed presteren als de patholoog zonder tijdslimiet en significant beter zijn dan een patholoog in een diagnostische situatie waarbij de tijd voor een casus altijd beperkt is.

De wisselwerking tussen een tumor en diens omgevende weefsel (het micromilieu) wordt gezien als een cruciale factor voor tumorgroei en de kans op metastasering. Dit micromilieu, en in het bijzonder de stromale cellen, ondergaan drastische veranderingen tijdens de groei van een tumor. In **Hoofdstuk 8** passen we kunstmatige intelligentie technieken toe om automatische dergelijke veranderingen in WSI te beoordelen. Eerst hebben we een deep learning-algoritme ontwikkeld dat in

staat is om een biopt te diagnosticeren (kanker versus normaal) enkel en alleen op basis van de visuele eigenschappen van het stroma. Vervolgens hebben we dit algoritme toegepast op DCIS laesies, om te bepalen in welke mate deze afwijkingen in staat zijn hun micromilieu dusdanig benvloeden dat er een omgeving ontstaat die tumorinvasie ondersteunt.

Publications

Papers in international journals

Ehteshami Bejnordi B, Veta M, van Diest P, van Ginneken B, Karssemeijer N, Litjens G, van der Laak J, The CAMELYON16 Consortium. "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer". *Accepted for publication in the Journal of the American Medical Association*, 2017.

Ehteshami Bejnordi B, Zuidhof G, Balkenhol M, Hermsen M, Bult P, van Ginneken B, Karssemeijer N, Litjens G, van der Laak J. "Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology image". *Accepted for publication in the Journal of Medical Imaging*, 2017.

Ehteshami Bejnordi B, Balkenhol M, Litjens G, Holland R, Bult P, Karssemeijer N, van der Laak J. "Automated detection of DCIS in whole-slide H&E stained breast histopathology images". *IEEE Transactions on Medical Imaging*, 35(9):2141-50, 2016.

Ehteshami Bejnordi B, Litjens G, Timofeeva N, Otte-Hller I, Homeyer A, Karssemeijer N, van der Laak J. "Stain specific standardization of whole-slide histopathological images". *IEEE Transactions on Medical Imaging*, 35(2):404-15, 2016.

Ehteshami Bejnordi B, Mullooly M, Pfeiffer R, Palakal M, Vacek P, Weaver D, Sprague B, Brinton L, van Ginneken B, Karssemeijer N, Beck A, Gierach G, van der Laak J, Sherman M. "Identification and characterization of tumor-associated stromal alterations using deep convolutional neural networks". *Under review*, 2017.

Mullooly M*, **Ehteshami Bejnordi B***, Pfeiffer R, Fan S, Palakal M, Hada M, Vacek P, Weaver D, Shepherd J, Fan B, Mahmoudzadeh A, Wang J, Johnson J, Herschorn S, Sprague B, Brinton L, Karssemeijer N, van der Laak J, Beck A, Sherman M, Gierach G. "Application of convolutional neural networks to breast biopsies to delineate the tissue correlates of mammographic breast density". *Under preparation*, 2017.

Litjens G, Kooi T, **Ehteshami Bejnordi B**, Setio AA, Ciompi F, Ghafoorian M, van der Laak JA, van Ginneken B, Snchez CI. "A survey on deep learning in medical image analysis"., *Medical Image Analysis*, 42:60-88, 2017.

Mertzanidou T, Hipwell JH, Reis S, Hawkes DJ, **Ehteshami Bejnordi B**, Dalmis M, Vreemann S, Platel B, van der Laak J, Karssemeijer N, Hermsen M, Bult P, Mann R. "3D volume reconstruction from serial breast specimen radiographs for mapping between histology and 3D whole specimen imaging". *Medical Physics*, 44(3):2473-4209, 2017.

Papers in conference proceedings

Ehteshami Bejnordi B, Linz J, Glass B, Mullooly M, Gierach GL, Sherman ME, Karssemeijer N, van der Laak J, Beck AH. "Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images". In: *proceedings of the IEEE International Symposium on Biomedical Imaging*, page 929-932, 2017.

Ehteshami Bejnordi B, Litjens G, Hermsen M, Karssemeijer N, van der Laak J. "A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images". In: *Medical Imaging*, volume 9420 of Proceedings of the SPIE, page 94200H, 2015.

Ehteshami Bejnordi B, Timofeeva N, Otte-Hller I, Karssemeijer N, van der Laak JA. "Quantitative analysis of stain variability in histology slides and an algorithm for standardization". In: *Medical Imaging*, volume 9041 of Proceedings of the SPIE, page 904108, 2014.

Litjens G, **Ehteshami Bejnordi B**, Timofeeva N, Swadi G, Kovacs I, Hulsbergen-van de Kaa C, van der Laak J. "Automated detection of prostate cancer in digitized whole-slide images of H and E-stained biopsy specimens". In: *Medical Imaging*, volume 9420 of Proceedings of the SPIE, page 94200B, 2015.

Ghazvinian Zanjani F, Zinger S, **Ehteshami Bejnordi B**, van der Laak J, Peter de With. "Stain Normalization of Histopathology Images Using Generative Adversarial Networks". *Submitted to the IEEE International Symposium on Biomedical Imaging*, 2018.

Ciompi F, Geessink O, **Ehteshami Bejnordi B**, de Souza GS, Baidoshvili A, Litjens G, van Ginneken B, Nagtegaal I, van der Laak J. "The importance of stain normalization in colorectal tissue classification with convolutional networks". In: *proceedings of the IEEE International Symposium on Biomedical Imaging*, page 160-163, 2017.

Mertzanidou T, Hipwell JH, Reis S, **Ehteshami Bejnordi B**, Hermsen M, Dalmis M, Vreemann S, Platel B, van der Laak J, Karssemeijer N, Mann R. "Whole Mastectomy Volume Reconstruction from 2D Radiographs and Its Mapping to Histology". In: *International Workshop on Digital Mammography*, page 367-374, 2016.

Van de Leemput S, Dorssers F, **Ehteshami Bejnordi B.**"Automatic detection of spiculation of pulmonary nodules in Computed Tomography images". In: *Medical Imaging*, volume 9414 of Proceedings of the SPIE, page 94142P, 2015.

Mahmood Q, Chodorowski A, **Ehteshami Bejnordi B**, Persson M. "A fully auto-

matic unsupervised segmentation framework for the brain tissues in MR images". In: *Medical Imaging*, volume 9038 of Proceedings of the SPIE, page 90381M, 2014.

Ehteshami Bejnordi B, Moshavegh R, Sujathan K, Malm P, Bengtsson E, Mehnert A. "Novel chromatin texture features for the classification of pap smears". In: *Medical Imaging*, volume 8676 of Proceedings of the SPIE, page 867608, 2013.

Moshavegh R, **Ehteshami Bejnordi B**, Mehnert A, Sujathan K, Malm P, Bengtsson E. "Automated segmentation of free-lying cell nuclei in Pap smears for malignancy-associated change analysis". In: *Engineering in Medicine and Biology Society (EMBC)*, page 5372-5375, 2012.

Bibliography

[1] Siegel R. L., Miller K. D., and Jemal A. Cancer statistics, 2017. *CA: A Cancer Journal for Clinicians*, 2016.

[2] Burstein H. J., Polyak K., Wong J. S., Lester S. C., and Kaelin C. M. Ductal carcinoma in situ of the breast. *New England Journal of Medicine*, 350(14):1430–1441, 2004.

[3] Schnitt S. J., Connolly J. L., Tavassoli F. A., Fechner R. E., Kempson R. L., Gelman R., and Page D. L. Interobserver reproducibility in the diagnosis of ductal proliferative breast lesions using standardized criteria. *The American journal of surgical pathology*, 16(12):1133–1143, 1992.

[4] Tuominen V. J. and Isola J. Linking whole-slide microscope images with dicom by using jpeg2000 interactive protocol. *Journal of Digital Imaging*, 23(4):454–462, 2010.

[5] Goode A., Gilbert B., Harkes J., Jukic D., and Satyanarayanan M. Openslide: A vendor-neutral software foundation for digital pathology. *Journal of pathology informatics*, 4, 2013.

[6] Wang Y., Williamson K. E., Kelly P. J., James J. A., and Hamilton P. W. Surfaceslide: A multitouch digital pathology platform. *PLOS ONE*, 7(1):1–12, 01 2012.

[7] Griffin J. and Treanor D. Digital pathology in clinical use: where are we now and what is holding us back? *Histopathology*, 70(1):134–145, 2017.

[8] Madabhushi A. and Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Medical Image Analysis*, 33:170 – 175, 2016.

[9] Meyer J. and Par G. Telepathology impacts and implementation challenges: A scoping review. *Archives of Pathology & Laboratory Medicine*, 139(12):1550–1557, 2015.

[10] McCarthy J., Minsky M. L., Rochester N., and Shannon C. E. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI magazine*, 27(4):12, 2006.

[11] Turing A. M. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[12] Newell A. and Simon H. The logic theory machine–a complex information processing system. *IRE Transactions on information theory*, 2(3):61–79, 1956.

[13] Samuel A. L. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 44(1.2):206–226, 2000.

[14] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[15] Jackson P. *Introduction to expert systems*. Addison-Wesley Pub. Co., Reading, MA, 1986.

[16] Gilbert F. J., Astley S. M., McGee M. A., Gillan M. G., Boggis C. R., Griffiths P. M., and Duffy S. W. Single reading with computer-aided detection and double reading of screening mammograms in the united kingdom national breast screening program. *Radiology*, 241(1):47–53, 2006.

[17] Khoo L. A., Taylor P., and Given-Wilson R. M. Computer-aided detection in the united kingdom national breast screening programme: prospective study. *Radiology*, 237(2):444–449, 2005.

[18] Murphy K., van Ginneken B., Schilham A. M., De Hoop B., Gietema H., and Prokop M. A large-scale evaluation of automatic pulmonary nodule detection in chest ct using local image features and k-nearest-neighbour classification. *Medical image analysis*, 13(5):757–770, 2009.

[19] Jacobs C., van Rikxoort E. M., Twellmann T., Scholten E. T., de Jong P. A., Kuhnigk J.-M., Oud-

kerk M., de Koning H. J., Prokop M., Schaefer-Prokop C., et al. Automatic detection of subsolid pulmonary nodules in thoracic computed tomography images. *Medical image analysis*, 18(2): 374–384, 2014.

[20] Akram M. U., Khalid S., Tariq A., Khan S. A., and Azam F. Detection and classification of retinal lesions for grading of diabetic retinopathy. *Computers in biology and medicine*, 45:161–171, 2014.

[21] Roychowdhury S., Koozekanani D. D., and Parhi K. K. Dream: diabetic retinopathy analysis using machine learning. *IEEE journal of biomedical and health informatics*, 18(5):1717–1728, 2014.

[22] Ghafoorian M., Karssemeijer N., van Uden I., de Leeuw F. E., Heskes T., Marchiori E., and Platel B. Small white matter lesion detection in cerebral small vessel disease. In *SPIE Medical Imaging*, pages 941411–941411. International Society for Optics and Photonics, 2015.

[23] Vijverberg K., Ghafoorian M., van Uden I. W., de Leeuw F.-E., Platel B., and Heskes T. A single-layer network unsupervised feature learning method for white matter hyperintensity segmentation. *SPIE Medical Imaging, International Society for Optics and Photonics*, pages 97851C–97851C, 2016.

[24] van den Heuvel T., van der Eerden A., Manniesing R., Ghafoorian M., Tan T., Andriessen T., Vyvere T. V., van den Hauwe L., Ter Haar Romeny B., Goraj B., et al. Automated detection of cerebral microbleeds in patients with traumatic brain injury. *NeuroImage: Clinical*, 12:241–251, 2016.

[25] Veta M., Pluim J. P. W., van Diest P. J., and Viergever M. A. Breast cancer histopathology image analysis: A review. *IEEE Transactions on Biomedical Engineering*, 61(5):1400–1411, May 2014.

[26] McCulloch W. S. and Pitts W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[27] Schmidhuber J. Deep learning in neural networks: An overview. *Neural Networks*, 61:85 – 117, 2015.

[28] Bengio Y. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127, January 2009.

[29] Eigen D., Rolfe J. T., Fergus R., and LeCun Y. Understanding deep architectures using a recursive convolutional network. *CoRR*, abs/1312.1847, 2013.

[30] Goodfellow I., Bengio Y., and Courville A. *Deep Learning*. MIT Press, 2016.

[31] Krizhevsky A., Sutskever I., and Hinton G. E. ImageNet classification with deep convolutional neural networks. In Pereira F., Burges C. J. C., Bottou L., and Weinberger K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[32] Cireşan D. C., Giusti A., Gambardella L. M., and Schmidhuber J. *Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks*, pages 411–418. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-40763-5.

[33] Ciresan D., Giusti A., Gambardella L. M., and Schmidhuber J. Deep neural networks segment neuronal membranes in electron microscopy images. In Pereira F., Burges C. J. C., Bottou L., and Weinberger K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 2843–2851. Curran Associates, Inc., 2012.

[34] Wang D., Khosla A., Gargeya R., Irshad H., and Beck A. H. Deep learning for identifying

metastatic breast cancer. *CoRR*, abs/1606.05718, 2016.

[35] Bejnordi B. E., Mitko V., Paul J. v. D., and et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22): 2199–2210, 2017.

[36] Litjens G. J. S., Kooi T., Bejnordi B. E., Setio A. A. A., Ciompi F., Ghafoorian M., van der Laak J. A. W. M., van Ginneken B., and Sánchez C. I. A survey on deep learning in medical image analysis. *CoRR*, abs/1702.05747, 2017.

[37] Ghaznavi F., Evans A., Madabhushi A., and Feldman M. Digital imaging in pathology: whole-slide imaging and beyond. *Annual Review of Pathology: Mechanisms of Disease*, 8:331–359, 2013.

[38] Ismail S. M., Colclough A. B., Dinnen J. S., Eakins D., Evans D., Gradwell E., O'sullivan J. P., Summerell J. M., and Newcombe R. G. Observer variation in histopathological diagnosis and grading of cervical intraepithelial neoplasia. *British Medical Journal*, 298(6675):707–710, 1989.

[39] Andrion A., Magnani C., Betta P. G., Donna A., Mollo F., Scelsi M., Bernardi P., Botta M., and Terracini B. Malignant mesothelioma of the pleura: interobserver variability. *Journal of clinical pathology*, 48(9):856–860, Sep 1995.

[40] Gurcan M. N., Boucheron L., Can A., Madabhushi A., Rajpoot N., and Yener B. Histopathological image analysis: A review. *IEEE Rev Biomed Eng*, 2:147–171, 2009.

[41] Niethammer M., Borland D., Marron J. S., Woosley J., and Thomas N. E. Appearance normalization of histology slides. *Mach Learn Med Imaging*, 6357:58–66, 2010.

[42] Ruifrok A. C., Johnston D. A., et al. Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology*, 23(4):291–299, 2001.

[43] Macenko M., Niethammer M., Marron J. S., Borland D., Woosley J. T., Guan X., Schmitt C., and Thomas N. E. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110, June 2009.

[44] Bağci U. and Bai L. Registration of standardized histological images in feature space. volume 6914, pages 69142V–69142V–9, 2008.

[45] Basavanhally A. and Madabhushi A. Em-based segmentation-driven color standardization of digitized histopathology. In *SPIE Medical Imaging*, pages 86760G–86760G. International Society for Optics and Photonics, 2013.

[46] van der Laak J. A., Pahlplatz M. M., Hanselaar A. G., and de Wilde P. C. Hue-saturation-density (hsd) model for stain recognition in digital images from transmitted light microscopy. *Cytometry*, 39(4):275–284, 2000.

[47] Fischer A. H., Jacobson K. A., Rose J., and Zeller R. Hematoxylin and eosin staining of tissue and cell sections. *Cold Spring Harbor Protocols*, 2008(5):pdb–prot4986, 2008.

[48] Bishop C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738.

[49] Rencher A. *Methods of Multivariate Analysis*. Wiley Series in Probability and Statistics. Wiley, 2003. ISBN 9780471461722.

[50] Nyul L. G., Udupa J. K., and Zhang X. New variants of a method of mri scale standardization. *IEEE Transactions on Medical Imaging*, 19(2):143–150, Feb 2000.

[51] Epstein J., Allsbrook Jr W., Amin M., and Egevad L. Update on the gleason grading system for prostate cancer: results of an international consensus conference of urologic pathologists. *Advances in anatomic pathology*, 13(1):57, 2006.

[52] Stoler M. H., Schiffman M., et al. Interobserver reproducibility of cervical cytologic and histologic interpretations: realistic estimates from the ascus-lsil triage study. *Jama*, 285(11):1500–1505, 2001.

[53] Roberts C. A., Beitsch P. D., Litz C. E., Hilton D. S., Ewing G. E., Clifford E., Taylor W., Hapke M. R., Babaian A., Khalid I., et al. Interpretive disparity among pathologists in breastsentinel lymph node evaluation. *The American journal of surgery*, 186(4):324–329, 2003.

[54] Bancroft J. D. and Gamble M. *Theory and practice of histological techniques*. Elsevier Health Sciences, 2008.

[55] Magee D., Treanor D., Crellin D., Shires M., Smith K., Mohee K., and Quirke P. Colour normalisation in digital histopathology images. In *Proceedings of Optical Tissue Image Analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*, pages 20–24, 2009.

[56] Khan A., Rajpoot N., Treanor D., and Magee D. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *Biomedical Engineering, IEEE Transactions on*, 61(6):1729–1738, June 2014.

[57] Reinhard E., Ashikhmin M., Gooch B., and Shirley P. Color transfer between images. *IEEE Comput. Graph. Appl.*, 21(5):34–41, September 2001.

[58] Ehteshami Bejnordi B., Timofeeva N., Otte-Höller I., Karssemeijer N., and van der Laak J. A. Quantitative analysis of stain variability in histology slides and an algorithm for standardization. In *SPIE Medical Imaging*, pages 904108–904108. International Society for Optics and Photonics, 2014.

[59] Bautista P. A., Hashimoto N., and Yagi Y. Color standardization in whole slide imaging using a color calibration slide. *Journal of pathology informatics*, 5, 2014.

[60] Hunter R. S. Accuracy, precision, and stability of new photoelectric color-difference meter. *Journal of the Optical Society of America*, 38(12):1094–1094, 1948.

[61] Homeyer A., Schenk A., Arlt J., Dahmen U., Dirsch O., and Hahn H. K. Practical quantification of necrosis in histological whole-slide images. *Computerized Medical Imaging and Graphics*, 37(4): 313–322, 2013.

[62] Litjens G. Automated slide analysis platform (asap). `https://github.com/GeertLitjens/ASAP`, 2017.

[63] Goode A., Gilbert B., Harkes J., Jukic D., and Satyanarayanan M. Openslide: A vendor-neutral software foundation for digital pathology. *Journal of pathology informatics*, 4, 2013.

[64] Cheng Z. and Liu Y. Efficient technique for ellipse detection using restricted randomized hough transform. In *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on*, volume 2, pages 714–718. IEEE, 2004.

[65] Loy G. and Zelinsky A. Fast radial symmetry for detecting points of interest. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(8):959–973, 2003.

[66] Moshavegh R., Bejnordi B., Mehnert A., Sujathan K., Malm P., and Bengtsson E. Automated

segmentation of free-lying cell nuclei in pap smears for malignancy-associated change analysis. In *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*, pages 5372–5375. IEEE, 2012.

[67] Peura M. and Iivarinen J. Efficiency of simple shape descriptors. In *Proceedings of the third international workshop on visual form*, volume 443, page 451. Citeseer, 1997.

[68] Ojala T., Pietikainen M., and Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[69] Zweig M. H. and Campbell G. Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine. *Clinical chemistry*, 39(4):561–577, 1993.

[70] Hanley J. A. and McNeil B. J. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843, 1983.

[71] Esgiar A. N., Naguib R. N. G., Sharif B. S., Bennett M. K., and Murray A. Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa. *IEEE Transactions on Information Technology in Biomedicine*, 2(3):197–203, Sept 1998.

[72] Tabesh A., Kumar V. P., Pang H.-Y., Verbel D., Kotsianti A., Teverovskiy M., and Saidi O. Automated prostate cancer diagnosis and gleason grading of tissue microarrays. In *Medical Imaging*, pages 58–70. International Society for Optics and Photonics, 2005.

[73] Viola P. and Jones M. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–511–I–518 vol.1, 2001.

[74] Doyle S., Madabhushi A., Feldman M., and Tomaszeweski J. A boosting cascade for automated detection of prostate cancer from digitized histology. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006*, pages 504–511. Springer, 2006.

[75] Freund Y. and Schapire R. E. Experiments with a new boosting algorithm. In *International Conference on Machine Learning*, pages 148–156. Morgan Kaufmann, 1996.

[76] Sertel O., Kong J., Shimada H., Catalyurek U., Saltz J. H., and Gurcan M. N. Computer-aided prognosis of neuroblastoma on whole-slide images: Classification of stromal development. *Pattern recognition*, 42(6):1093–1103, 2009.

[77] Ji S., Wei B., Yu Z., Yang G., and Yin Y. A new multistage medical segmentation method based on superpixel and fuzzy clustering. *Computational and mathematical methods in medicine*, 2014, 2014.

[78] Gorelick L., Veksler O., Gaed M., Gmez J. A., Moussa M., Bauman G., Fenster A., and Ward A. D. Prostate histopathology: Learning tissue component histograms for cancer detection and classification. *IEEE Transactions on Medical Imaging*, 32(10):1804–1818, Oct 2013.

[79] Achanta R., Shaji A., Smith K., Lucchi A., Fua P., and Susstrunk S. Slic superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2274–2282, 2012.

[80] Beck A. H., Sangoi A. R., Leung S., Marinelli R. J., Nielsen T. O., van de Vijver M. J., West R. B.,

van de Rijn M., and Koller D. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine*, 3(108):108ra113–108ra113, 2011.

[81] Duan K.-B., Rajapakse J. C., Wang H., and Azuaje F. Multiple svm-rfe for gene selection in cancer classification with expression data. *IEEE Transactions on NanoBioscience*, 4(3):228–234, Sept 2005.

[82] Deng H. and Runger G. Gene selection with guided regularized random forest. *Pattern Recognition*, 46(12):3483 – 3489, 2013.

[83] Siegel R. L., Miller K. D., and Jemal A. Cancer statistics, 2015. *CA: a cancer journal for clinicians*, 65(1):5–29, 2015.

[84] Ellis I. O. Intraductal proliferative lesions of the breast: morphology, associated risk and molecular biology. *Modern pathology*, 23:S1–S7, 2010.

[85] Simpson P. T., Reis-Filho J. S., Gale T., and Lakhani S. R. Molecular evolution of breast cancer. *The Journal of pathology*, 205(2):248–254, 2005.

[86] Dong F., Irshad H., Oh E.-Y., Lerwill M. F., Brachtel E. F., Jones N. C., Knoblauch N. W., Montaser-Kouhsari L., Johnson N. B., Rao L. K. F., Faulkner-Jones B., Wilbur D. C., Schnitt S. J., and Beck A. H. Computational pathology to discriminate benign from malignant intraductal proliferations of the breast. *PLoS ONE*, 9(12):e114885, 12 2014.

[87] Dundar M. M., Badve S., Bilgin G., Raykar V., Jain R., Sertel O., and Gurcan M. N. Computerized classification of intraductal breast lesions using histopathological images. *Biomedical Engineering, IEEE Transactions on*, 58(7):1977–1984, 2011.

[88] Srinivas U., Mousavi H., Jeon C., Monga V., Hattel A., and Jayarao B. Shirc: A simultaneous sparsity model for histopathological image representation and classification. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pages 1118–1121. IEEE, 2013.

[89] Peikari M., Gangeh M., Zubovits J., Clarke G., and Martel A. Triaging diagnostically relevant regions from pathology whole slides of breast cancer: A texture based approach. *Medical Imaging, IEEE Transactions on*, 35(1):307 – 315, 2015.

[90] Bejnordi B. E., Litjens G., Hermsen M., Karssemeijer N., and van der Laak J. A. A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images. In *SPIE Medical Imaging*, pages 94200H–94200H. International Society for Optics and Photonics, 2015.

[91] Jaromczyk J. and Toussaint G. Relative neighborhood graphs and their relatives. *Proceedings of the IEEE*, 80(9):1502–1517, Sep 1992.

[92] Moore E. F. The shortest path through a maze. In *Proceedings of an International Symposium on the Theory of Switching*, pages 285–292. Harvard University Press, 1957.

[93] Xu R., Wunsch D., et al. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678, 2005.

[94] Schaeffer S. E. Graph clustering. *Computer Science Review*, 1(1):27–64, 2007.

[95] Dörrie H. *100 great problems of elementary mathematics*. Courier Corporation, 2013.

[96] Ehteshami Bejnordi B., Litjens G., Timofeeva N., Otte-Holler I., Homeyer A., Karssemeijer N.,

and van der Laak J. Stain specific standardization of whole-slide histopathological images. *Medical Imaging, IEEE Transactions on*, 35(2):404–415, 2015.

[97] Haralick R. M., Shanmugam K., and Dinstein I. H. Textural features for image classification. *Systems, Man and Cybernetics, IEEE Transactions on*, (6):610–621, 1973.

[98] Topi M., Timo O., Matti P., and Maricor S. Robust texture classification by subsets of local binary patterns. In *Pattern Recognition, 2000. Proceedings. 15th International Conference on*, volume 3, pages 935–938. IEEE, 2000.

[99] Doyle S., Agner S., Madabhushi A., Feldman M., and Tomaszewski J. Automated grading of breast cancer histopathology using spectral clustering with textural and architectural image features. In *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pages 496–499. IEEE, 2008.

[100] Naik S., Doyle S., Agner S., Madabhushi A., Feldman M., and Tomaszewski J. Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology. In *Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on*, pages 284–287, May 2008.

[101] Friedman J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 29:1189–1232, 2001.

[102] Pinder S. E. and Ellis I. O. Ductal carcinoma in situ (dcis) and atypical ductal hyperplasia (adh)-current definitions and classification. *Breast Cancer Research*, 5(5):254–257, 2003.

[103] Page D. Atypical hyperplastic lesions of the female breast: A long-term follow-up study. *Plastic and Reconstructive Surgery*, 77(4):688, 1986.

[104] Page D. L. and Rogers L. W. Combined histologic and cytologic criteria for the diagnosis of mammary atypical ductal hyperplasia. *Human pathology*, 23(10):1095–1097, 1992.

[105] Tavassoli F. Intraduct hyperplasias, ordinary and atypical. *Pathology of the Breast*, pages 155–191, 1992.

[106] Bunch P. C., Hamilton J. F., Sanderson G. K., and Simmons A. H. A free response approach to the measurement and characterization of radiographic observer performance. In *Application of Optical Instrumentation in Medicine VI*, pages 124–135. International Society for Optics and Photonics, 1977.

[107] Nguyen K., Sarkar A., and Jain A. K. Prostate cancer grading: Use of graph cut and spatial arrangement of nuclei. *Medical Imaging, IEEE Transactions on*, 33(12):2254–2270, 2014.

[108] Naik S., Doyle S., Feldman M., Tomaszewski J., and Madabhushi A. Gland segmentation and computerized gleason grading of prostate histology by integrating low-, high-level and domain specific information. In *MIAAB workshop*, pages 1–8. Citeseer, 2007.

[109] Gunduz-Demir C., Kandemir M., Tosun A. B., and Sokmensuer C. Automatic segmentation of colon glands using object-graphs. *Medical image analysis*, 14(1):1–12, 2010.

[110] Fakhrzadeh A., Sporndly-Nees E., Holm L., and Hendriks C. L. L. Analyzing tubular tissue in histopathological thin sections. In *Digital Image Computing Techniques and Applications (DICTA), 2012 International Conference on*, pages 1–6. IEEE, 2012.

[111] Guray M. and Sahin A. A. Benign breast diseases: classification, diagnosis, and management.

The Oncologist, 11(5):435–449, 2006.

[112] Sirinukunwattana K., Snead D. R., and Rajpoot N. M. A stochastic polygons model for glandular structures in colon histology images. *Medical Imaging, IEEE Transactions on*, 34(11):2366–2378, 2015.

[113] Dupont W. D., Parl F. F., Hartmann W. H., Brinton L. A., Winfield A. C., Worrell J. A., Schuyler P. A., and Plummer W. D. Breast cancer risk associated with proliferative breast disease and atypical hyperplasia. *CANCER-PHILADELPHIA-*, 71:1258–1258, 1993.

[114] Ehteshami Bejnordi B., Balkenhol M., Litjens G., Holland R., Bult P., Karssemeijer N., and van der Laak J. Automated detection of DCIS in whole-slide H&E stained breast histopathology images. *IEEE Transactions on Medical Imaging*, 35(9):2141–2150, Sept 2016.

[115] Balazsi M., Blanco P., Zoroquiain P., Levine M. D., and Burnier, Jr. M. N. Invasive ductal breast carcinoma detector that is robust to image magnification in whole digital slides. *Journal of Medical Imaging*, 3(2):027501, 2016.

[116] Cruz-Roa A., Basavanhally A., and González et al. F. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In *SPIE medical imaging*, pages 904103–904103. International Society for Optics and Photonics, 2014.

[117] Cruz-Roa A., Gilmore H., Basavanhally A., Feldman M., Ganesan S., Shih N. N., Tomaszewski J., González F. A., and Madabhushi A. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. *Scientific Reports*, 7, 2017.

[118] Rezaeilouyeh H., Mollahosseini A., and Mahoor M. H. Microscopic medical image classification framework via deep learning and shearlet transform. *Journal of Medical Imaging*, 3(4): 044501, 2016.

[119] Ehteshami Bejnordi B., Lin J., Glass B., Mullooly M., Gierach G. L., Sherman M. E., Karssemeijer N., van der Laak J., and Beck A. H. Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images. pages 929–932, April 2017.

[120] Ehteshami Bejnordi B., Litjens G., Hermsen M., Karssemeijer N., and van der Laak J. A. A multi-scale superpixel classification approach to the detection of regions of interest in whole slide histopathology images. In *SPIE Medical Imaging*, pages 94200H–94200H. International Society for Optics and Photonics, 2015.

[121] Breiman L. Random forests. *Machine learning*, 45(1):5–32, 2001.

[122] Gu J., Wang Z., Kuen J., Ma L., Shahroudy A., Shuai B., Liu T., Wang X., and Wang G. Recent advances in convolutional neural networks. *arXiv:1512.07108*, 2015.

[123] LeCun Y., Bengio Y., and Hinton G. Deep learning. *Nature*, 521(7553):436–444, 2015.

[124] Hinton G., Deng L., Yu D., Dahl G. E., Mohamed A.-r., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T. N., et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

[125] Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., Berg A. C., and Fei-Fei L. ImageNet large scale visual recognition challenge.

IJCV, 115(3):1–42, 2014.

[126] Silver D., Huang A., Maddison C. J., Guez A., Sifre L., Van Den Driessche G., Schrittwieser J., Antonoglou I., Panneershelvam V., Lanctot M., et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

[127] Gulshan V., Peng L., Coram M., Stumpe M. C., Wu D., Narayanaswamy A., Venugopalan S., Widner K., Madams T., Cuadros J., et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22): 2402–2410, 2016.

[128] Esteva A., Kuprel B., Novoa R. A., Ko J., Swetter S. M., Blau H. M., and Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

[129] Ehteshami Bejnordi B. and van der Laak J. Camelyon16: Grand challenge on cancer metastasis detection in lymph nodes 2016. `https://camelyon16.grand-challenge.org`.

[130] Litjens G., Sánchez C. I., Timofeeva N., Hermsen M., Nagtegaal I., Kovacs I., Hulsbergen-van de Kaa C., Bult P., van Ginneken B., and van der Laak J. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports*, 6, 2016.

[131] Mehrtash A., Sedghi A., Ghafoorian M., Taghipour M., Tempany C. M., Wells III W. M., Kapur T., Mousavi P., Abolmaesumi P., and Fedorov A. Classification of clinical significance of mri prostate findings using 3d convolutional neural networks. In *Proceedings of SPIE–the International Society for Optical Engineering*, volume 10134. NIH Public Access, 2017.

[132] Ghafoorian M. and Platel B. Convolutional neural networks for ms lesion segmentation, method description of diag team. *Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge*, pages 1–2, 2015.

[133] He K., Zhang X., Ren S., and Sun J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[134] Zagoruyko S. and Komodakis N. Wide residual networks. *CoRR*, abs/1605.07146, 2016.

[135] Simonyan K. and Zisserman A. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[136] Ehteshami Bejnordi B., Moshavegh R., Sujathan K., Malm P., Bengtsson E., and Mehnert A. Novel chromatin texture features for the classification of pap smears. In *SPIE Medical Imaging*, pages 867608–867608. International Society for Optics and Photonics, 2013.

[137] He K., Zhang X., Ren S., and Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[138] Cohen J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[139] Ghafoorian M., Karssemeijer N., Heskes T., van Uder I., de Leeuw F.-E., Marchiori E., van Ginneken B., and Platel B. Non-uniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pages 1414–1417. IEEE, 2016.

[140] Ghafoorian M., Karssemeijer N., Heskes T., Bergkamp M., Wissink J., Obels J., Keizer K.,

de Leeuw F.-E., van Ginneken B., Marchiori E., et al. Deep multi-scale location-aware 3d convolutional neural networks for automated detection of lacunes of presumed vascular origin. *NeuroImage: Clinical*, 14:391–399, 2017.

[141] Ghafoorian M., Karssemeijer N., Heskes T., van Uden I., Sanchez C., Litjens G., de Leeuw F.-E., van Ginneken B., Marchiori E., and Platel B. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *arXiv preprint arXiv:1610.04834*, 2016.

[142] Ronneberger O., Fischer P., and Brox T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[143] Dalmı ş M. U., Litjens G., Holland K., Setio A., Mann R., Karssemeijer N., and Gubern-Mrida A. Using deep learning to segment breast and fibroglandular tissue in mri volumes. *Medical Physics*, 44(2):533–546, 2017.

[144] Foschini M. P., Scarpellini F., Gown A. M., and Eusebi V. Differential expression of myoepithelial markers in salivary, sweat and mammary glands. *International journal of surgical pathology*, 8(1):29–37, 2000.

[145] Ciompi F., Geessink O., Ehteshami Bejnordi B., de Souza G. S., Baidoshvili A., Litjens G., van Ginneken B., Nagtegaal I., and van der Laak J. The importance of stain normalization in colorectal tissue classification with convolutional networks. *arXiv preprint arXiv:1702.05931*, 2017.

[146] Veta M. Tupac16: Tumor proliferation assessment challenge 2016. `http://tupac.tue-image.nl`.

[147] Farabet C., Couprie C., Najman L., and LeCun Y. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1915–1929, Aug 2013.

[148] Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., and Rabinovich A. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.

[149] Mikolov T., Deoras A., Povey D., Burget L., and Černockỳ J. Strategies for training large scale neural network language models. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 196–201. IEEE, 2011.

[150] Sainath T. N., Mohamed A.-r., Kingsbury B., and Ramabhadran B. Deep convolutional neural networks for lvcsr. In *Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on*, pages 8614–8618. IEEE, 2013.

[151] Graves A. and Jaitly N. Towards end-to-end speech recognition with recurrent neural networks. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages II–1764–II–1772. JMLR.org, 2014.

[152] Ma J., Sheridan R. P., Liaw A., Dahl G. E., and Svetnik V. Deep neural nets as a method for quantitative structureactivity relationships. *Journal of Chemical Information and Modeling*, 55(2): 263–274, 2015.

[153] Zhou J. and Troyanskaya O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Meth*, 12(10):931–934, Oct 2015.

[154] Xiong H. Y., Alipanahi B., Lee L. J., Bretschneider H., Merico D., Yuen R. K. C., Hua Y., Gueroussov S., Najafabadi H. S., Hughes T. R., Morris Q., Barash Y., Krainer A. R., Jojic N., Scherer S. W., Blencowe B. J., and Frey B. J. The human splicing code reveals new insights into the genetic determinants of disease. *Science*, 347(6218), 2015.

[155] Edge S. B. and Compton C. C. The american joint committee on cancer: the 7th edition of the ajcc cancer staging manual and the future of tnm. *Annals of surgical oncology*, 17(6):1471–1474, 2010.

[156] Vestjens J., Pepels M., de Boer M., Borm G. F., van Deurzen C. H., van Diest P. J., van Dijck J., Adang E., Nortier J. W., Rutgers E. T., et al. Relevant impact of central pathology review on nodal classification in individual breast cancer patients. *Annals of oncology*, 23(10):2561–2566, 2012.

[157] de Boer M., van Deurzen C. H., van Dijck J. A., Borm G. F., van Diest P. J., Adang E. M., Nortier J. W., Rutgers E. J., Seynaeve C., Menke-Pluymers M. B., et al. Micrometastases or isolated tumor cells and the outcome of breast cancer. *New England Journal of Medicine*, 361(7):653–663, 2009.

[158] Jatoi I., Hilsenbeck S. G., Clark G. M., and Osborne C. K. Significance of axillary lymph node metastasis in primary breast cancer. *Journal of clinical oncology*, 17(8):2334–2334, 1999.

[159] Meijering E., Carpenter A. E., Peng H., Hamprecht F. A., and Olivo-Marin J.-C. Imagining the future of bioimage analysis. *Nature Biotechnology*, 34(12):1250–1255, 2016.

[160] Reed J., Rosman M., Verbanac K. M., Mannie A., Cheng Z., and Tafra L. Prognostic implications of isolated tumor cells and micrometastases in sentinel nodes of patients with invasive breast cancer: 10-year analysis of patients enrolled in the prospective east carolina university/anne arundel medical center sentinel node multicenter study. *Journal of the American College of Surgeons*, 208(3):333–340, 2009.

[161] Chagpar A., Middleton L. P., Sahin A. A., Meric-Bernstam F., Kuerer H. M., Feig B. W., Ross M. I., Ames F. C., Singletary S. E., Buchholz T. A., et al. Clinical outcome of patients with lymph node-negative breast carcinoma who have sentinel lymph node micrometastases detected by immunohistochemistry. *Cancer*, 103(8):1581–1586, 2005.

[162] Mariani G., Moresco L., Viale G., Villa G., Bagnasco M., Canavese G., Buscombe J., Strauss H. W., and Paganelli G. Radioguided sentinel lymph node biopsy in breast cancer surgery. *Journal of Nuclear Medicine*, 42(8):1198–1215, 2001.

[163] Chakraborty D. P. and Winter L. Free-response methodology: alternate analysis and a new observer-performance experiment. *Radiology*, 174(3):873–881, 1990.

[164] Efron B. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pages 569–593. Springer, 1992.

[165] Obuchowski N. A., Beiden S. V., Berbaum K. S., Hillis S. L., Ishwaran H., Song H. H., and Wagner R. F. Multireader, multicase receiver operating characteristic analysis:: an empirical comparison of five methods1. *Academic radiology*, 11(9):980–995, 2004.

[166] Dorfman D. D., Berbaum K. S., and Metz C. E. Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method. *Investigative radiology*, 27(9):723–731, 1992.

[167] Hillis S. L., Obuchowski N. A., and Berbaum K. S. Power estimation for multireader roc methods: an updated and unified approach. *Academic radiology*, 18(2):129–142, 2011.

[168] Mason S. J. and Graham N. E. Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, 128(584):2145–2166, 2002.

[169] Lowe D. G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[170] Cortes C. and Vapnik V. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[171] Cserni G., Bianchi S., Boecker W., Decker T., Lacerda M., Rank F., and Wells C. A. Improving the reproducibility of diagnosing micrometastases and isolated tumor cells. *Cancer*, 103(2):358–367, 2005.

[172] Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., and Zitnick C. L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[173] Ghafoorian M., Mehrtash A., Kapur T., Karssemeijer N., Marchiori E., Pesteie M., Guttmann C. R., de Leeuw F.-E., Tempany C. M., van Ginneken B., et al. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. *arXiv preprint arXiv:1702.07841*, 2017.

[174] Bengio Y. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning Workshop - Volume 27*, UTLW'11, pages 17–37. JMLR.org, 2011.

[175] Yosinski J., Clune J., Bengio Y., and Lipson H. How transferable are features in deep neural networks? In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 3320–3328, Cambridge, MA, USA, 2014. MIT Press.

[176] Razavian A. S., Azizpour H., Sullivan J., and Carlsson S. Cnn features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, CVPRW '14, pages 512–519, Washington, DC, USA, 2014. IEEE Computer Society. ISBN 978-1-4799-4308-1.

[177] Chen L.-C., Papandreou G., Kokkinos I., Murphy K., and Yuille A. L. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015.

[178] Quail D. F. and Joyce J. A. Microenvironmental regulation of tumor progression and metastasis. *Nature medicine*, 19(11):1423–1437, 2013.

[179] Egeblad M., Nakasone E. S., and Werb Z. Tumors as organs: complex tissues that interface with the entire organism. *Developmental cell*, 18(6):884–901, 2010.

[180] Joyce J. A. and Pollard J. W. Microenvironmental regulation of metastasis. *Nature Reviews Cancer*, 9(4):239–252, 2009.

[181] Provenzano P. P., Eliceiri K. W., Campbell J. M., Inman D. R., White J. G., and Keely P. J. Collagen reorganization at the tumor-stromal interface facilitates local invasion. *BMC medicine*, 4(1):38, 2006.

[182] Orimo A., Gupta P. B., Sgroi D. C., Arenzana-Seisdedos F., Delaunay T., Naeem R., Carey V. J.,

Richardson A. L., and Weinberg R. A. Stromal fibroblasts present in invasive human breast carcinomas promote tumor growth and angiogenesis through elevated sdf-1/cxcl12 secretion. *Cell*, 121(3):335 – 348, 2005.

[183] Rønnov-Jessen L., Petersen O. W., Koteliansky V. E., and Bissell M. J. The origin of the myofibroblasts in breast cancer. recapitulation of tumor environment in culture unravels diversity and implicates converted fibroblasts and recruited smooth muscle cells. *Journal of Clinical Investigation*, 95(2):859, 1995.

[184] Tlsty T. D. and Hein P. W. Know thy neighbor: stromal cells can contribute oncogenic signals. *Current opinion in genetics & development*, 11(1):54–59, 2001.

[185] Ma X.-J., Dahiya S., Richardson E., Erlander M., and Sgroi D. C. Gene expression profiling of the tumor microenvironment during breast cancer progression. *Breast Cancer Research*, 11(1): R7, 2009.

[186] Lopez-Garcia M. A., Geyer F. C., Lacroix-Triki M., Marchi C., and Reis-Filho J. S. Breast cancer precursors revisited: molecular features and progression pathways. *Histopathology*, 57(2):171–192, 2010.

[187] Salgado R., Denkert C., Demaria S., Sirtaine N., Klauschen F., Pruneri G., Wienert S., Van den Eynden G., Baehner F. L., Penault-Llorca F., et al. The evaluation of tumor-infiltrating lymphocytes (tils) in breast cancer: recommendations by an international tils working group 2014. *Annals of oncology*, 26(2):259–271, 2014.

[188] Gierach et al. G. L. Comparison of mammographic density assessed as volumes and areas among women undergoing diagnostic image-guided breast biopsy. *Cancer Epidemiology and Prevention Biomarkers*, 23(11):1055–9965, 2014.

[189] Gierach G. L., Patel D. A., Pfeiffer R. M., Figueroa J. D., Linville L., Papathomas D., Johnson J. M., Chicoine R. E., Herschorn S. D., Shepherd J. A., Wang J., Malkov S., Vacek P. M., Weaver D. L., Fan B., Mahmoudzadeh A. P., Palakal M., Xiang J., Oh H., Horne H. N., Sprague B. L., Hewitt S. M., Brinton L. A., and Sherman M. E. Relationship of terminal duct lobular unit involution of the breast with area and volume mammographic densities. *Cancer Prevention Research*, 9(2):149–158, 2016.

[190] Pinder S. E. Ductal carcinoma in situ (dcis): pathological features, differential diagnosis, prognostic factors and specimen evaluation. *Modern Pathology*, 23:S8–S13, 2010.

[191] Bejnordi B. E., Linz J., Glass B., Mullooly M., Gierach G. L., Sherman M. E., Karssemeijer N., van der Laak J., and Beck A. H. Deep learning-based assessment of tumor-associated stroma for diagnosing breast cancer in histopathology images. *arXiv preprint arXiv:1702.05803*, 2017.

[192] Okabe A., Boots B., Sugihara K., and Chiu S. N. *Spatial tessellations: concepts and applications of Voronoi diagrams*, volume 501. John Wiley & Sons, 2009.

[193] Robin X., Turck N., Hainard A., Tiberti N., Lisacek F., Sanchez J.-C., and Müller M. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, 12 (1):77, 2011.

[194] Bejnordi B. E., Zuidhof G. C. A., Balkenhol M., Hermsen M., Bult P., van Ginneken B., Karssemeijer N., Litjens G. J. S., and van der Laak J. Context-aware stacked convolutional neural networks for classification of breast carcinomas in whole-slide histopathology images. *CoRR*,

abs/1705.03678, 2017.

[195] Solin L. J., Kurtz J., Fourquet A., Amalric R., Recht A., Bornstein B. A., Kuske R., Taylor M., Barrett W., Fowble B., et al. Fifteen-year results of breast-conserving surgery and definitive breast irradiation for the treatment of ductal carcinoma in situ of the breast. *Journal of clinical oncology*, 14(3):754–763, 1996.

[196] Silverstein M. J., Lagios M. D., Groshen S., Waisman J. R., Lewinsky B. S., Martino S., Gamagami P., and Colburn W. J. The influence of margin width on local control of ductal carcinoma in situ of the breast. *New England Journal of Medicine*, 340(19):1455–1461, 1999.

[197] Benson J. R., Jatoi I., and Toi M. Treatment of low-risk ductal carcinoma in situ: is nothing better than something? *The Lancet Oncology*, 17(10):e442–e451, 2016.

[198] Wallis M., Clements K., Kearins O., Ball G., Macartney J., and Lawrence G. The effect of dcis grade on rate, type and time to recurrence after 15 years of follow-up of screen-detected dcis. *British journal of cancer*, 106(10):1611–1617, 2012.

[199] Lagios M. D., Margolin F. R., Westdahl P. R., and Rose M. R. Mammographically detected duct carcinoma in situ. frequency of local recurrence following tylectomy and prognostic effect of nuclear grade on local recurrence. *Cancer*, 63(4):618–624, 1989.

[200] Bagnall M. J., Evans A. J., Wilson A. R. M., Pinder S. E., Denley H., Geraghty J. G., and Ellis I. O. Predicting invasion in mammographically detected microcalcification. *Clinical radiology*, 56(10):828–832, 2001.

[201] Bombonati A. and Sgroi D. C. The molecular pathology of breast cancer progression. *The Journal of pathology*, 223(2):308–318, 2011.

[202] Francis A., Bartlett J., Billingham L., Bowden S., Brookes C., Dodwell D., Evans A., Fallowfield L., Gaunt C., Hanby A., et al. Abstract ot2-3-01: The loris trial: A multicentre, randomized phase iii trial of standard surgery versus active monitoring in women with newly diagnosed low risk ductal carcinoma in situ, 2013.

[203] Elshof L. E., Tryfonidis K., Slaets L., van Leeuwen-Stok A. E., Skinner V. P., Dif N., Pijnappel R. M., Bijker N., Emiel J. T., Wesseling J., et al. Feasibility of a prospective, randomised, open-label, international multicentre, phase iii, non-inferiority trial to assess the safety of active surveillance for low risk ductal carcinoma in situ–the lord study. *European Journal of Cancer*, 51(12):1497–1510, 2015.

[204] Youngwirth L. M., Boughey J. C., and Hwang E. S. Surgery versus monitoring and endocrine therapy for low-risk dcis: the comet trial. *Cancer Inst*, 105:701–710, 2013.

[205] Lin M., Chen Q., and Yan S. Network in network. *CoRR*, abs/1312.4400, 2013.

[206] Bándi P., van de Loo R., Intezar M., Geijs D., Ciompi F., van Ginneken B., van der Laak J., and Litjens G. Comparison of different methods for tissue segmentation in histopathological whole-slide images. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 591–595, April 2017.

[207] Nesterov Y. A method of solving a convex programming problem with convergence rate O(1/sqr(k)). *Soviet Mathematics Doklady*, 27:372–376, 1983.

[208] Ioffe S. and Szegedy C. Batch normalization: Accelerating deep network training by reducing

internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[209] Nair V. and Hinton G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[210] Srivastava N., Hinton G. E., Krizhevsky A., Sutskever I., and Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[211] van Grinsven M. J. J. P., van Ginneken B., Hoyng C. B., Theelen T., and Snchez C. I. Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images. *IEEE Transactions on Medical Imaging*, 35(5):1273–1284, May 2016.

[212] Ben-David S., Blitzer J., Crammer K., and Pereira F. Analysis of representations for domain adaptation. In Schölkopf P. B., Platt J. C., and Hoffman T., editors, *Advances in Neural Information Processing Systems 19*, pages 137–144. MIT Press, 2007.

[213] Ben-David S., Blitzer J., Crammer K., Kulesza A., Pereira F., and Vaughan J. W. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

[214] Ganin Y., Ustinova E., Ajakan H., Germain P., Larochelle H., Laviolette F., Marchand M., and Lempitsky V. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

[215] Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., and Bengio Y. Generative adversarial nets. In Ghahramani Z., Welling M., Cortes C., Lawrence N. D., and Weinberger K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[216] Radford A., Metz L., and Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015.

[217] Denton E. L., Chintala S., szlam a., and Fergus R. Deep generative image models using a laplacian pyramid of adversarial networks. In Cortes C., Lawrence N. D., Lee D. D., Sugiyama M., and Garnett R., editors, *Advances in Neural Information Processing Systems 28*, pages 1486–1494. Curran Associates, Inc., 2015.

[218] Donahue J., Krähenbühl P., and Darrell T. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

Acknowledgments

Many people have contributed to the four wonderful years of my experience as a PhD student and for that I want to express my special thanks.

First, I must thank my adviser Jeroen van der Laak for his friendship, expertise, and long-standing support. It was an honor for me to be his first direct PhD student and my experience working with him has been nothing short of amazing. Jeroen created an invaluable space for me to do this research and develop myself as a researcher. I appreciate the freedom that was given to me to find my own research path along with all the guidance and support. He turned several of my messy paper drafts into well-written and well-organized manuscripts. Jeroen, the joy and enthusiasm you have for your research was contagious and motivational for me, even during tough times in my PhD pursuit. Thank you for this wonderful experience.

I am very grateful to my brilliant promoter, Nico Karssemeijer, for teaching me not just about several aspects of computer-aided diagnosis, machine learning and image analysis, but for teaching me how to think. Nico, I vividly remember how in the first few years of my PhD, you molded me from a confused student with many scattered ideas to a competent researcher more aware of the wider context. Your mix of straightforward criticism combined with heart-warming support and encouragement have given me great confidence as a researcher.

I would like to thank Geert Litjens, who overtime, I have come to regard as my adviser. He gave me his time and lent his expertise from the very beginning of my research. I highly appreciate Geert's support and his insightful comments and suggestions for my papers. I have a lot of respect and admiration for Geert and I am confident that in the near future he will be regarded as one of the main influential scientists in the field of computational pathology.

Many thanks to Bram van Ginneken for his help, advise, and comments on chapters 6, 7, and 8. I appreciate all the contribution of time, ideas, and the valuable feedback from Bram that helped me in the organization of the CAMELYON16 challenge. I would also like to acknowledge the deep learning group in DIAG which owes a debt of gratitude to Bram's vision and organizational persistence.

I also want to take the chance to thank my manuscript and PhD defense committee members. Tom Heskes, Nasir Rajpoot, and Jelle Wesseling, I feel proud and honored that you have accepted to be on my examination committee.

I would like to thank the reviewers and the editorial teams of JAMA and TMI for their valuable comments that significantly improved my papers. I was astounded by the level of professionalism of the JAMA editors, and the expertise and quality of comments from their reviewers. I do not gratefully thank the reviewers of Nature Medicine for their poor and useless comments.

Next, I would like to thank some people of outstanding importance for my re-

search. When I started my PhD in the digital pathology team, we were a small team of 5 members including Jeroen van der Laak, Irene Otte-Holler, Meyke Hermsen, Nadya Timofeeva, and myself. Irene, has contributed immensely to my first paper on stain standardization. Irene, from the very first few days, you educated me with scanning histopathology slides and introduced me to all the technical aspects of tissue preparation. Meyke had numerous contributions in many of my papers. Meyke, apart from your major contributions in data management, I learned a lot about many different aspects of breast and lymph node pathology from you. I also need to thank Peter Bult for his terrific support and ideas that gave me insight into the world of breast pathology. Peter was almost always available for immediate short meetings. I learned a lot from his insightful comments, feedback, and advice. Nadya, you've been a great pleasure to work with and also a wonderful friend. You were the first person in DIAG I had a technical meeting with. I clearly remember the first time we met and you described your project details beside the coffee machine of Radiology. How quickly time flies! The team of computational pathology has ever since grown and acquired an international presence. I will get back to thank the people who contributed to my research from this team in a couple of paragraphs.

I have to thank many people who contributed to my day-to-day life at the DIAG group and made my experience at Radboud so pleasant. DIAG has been a great source of friendship as well as thoughtful discussions and interactions. I should start with Mohsen, who was my best friend during this entire period. Mohsen, you have not only been a generator of novel perspectives and insightful comments in our discussions, but also an unlimited source of encouragement and support for me. I will forever be grateful to you and wish you all the greatest success and happiness. I am also thankful to many of the active members of DIAG meetings including DIAG discussion hour and deep learning meetings who have provoked thoughtful discussions and interactions. Specifically I want to thank Rashindra, Bram Platel, Clarisa, Henkjan, Francesco, Thijs, Colin, Arnaud, Jan Jurre, David, etc. Bram Platel, you have been a great friend and colleague. It was a pleasure collaborating with you and Mehmet on the VPHPRISM project. So many people in DIAG helped with friendship, inspiration, encouragement, or in other ways during my Ph.D. work. Thanks are due to them all: The fantastic Spanish speaking colleagues, Clarisa, Francesco (super multilingual), Albert, Leti, Alejandro, Jaime, Gabriel. Thanks for all the joyful moments, parties, games, and drinks. Albert, the fun we had at your PhD defence party is unbeatable. Many members of the breast group, Mehmet, Jan-Jurre, Suzan, Katharina, Tao, Jan van Zelst, Jonas Teuwen, and Christiana Balta. I had the privilege of being office-mates with many of you. Working at DIAG offices was hardly ever boring thanks to the following roomies (and, of course, myself): Mohsen, Jan-

Jurre, Jonas (mathematician/politician/computer scientist - How?), Mehmet, Pragnya, Katharina, Ajay, Sil, Midas, Thomas, etc. I want to extend my thanks to many members of the Lung group, Colin, Kaman, Jean-Paul, Arnaud, Rick, Steven, Sarah, Sjoerd. Jean-Paul, it is wonderful that we are defending/have defended in the same week. It has always been so nice talking to you and congrats for all the exciting good news happening to you (buying a house, defending your Ph.D. at the same time as mine, turning 30 next year (oh you are getting old!)). Special thanks to Pragnya for her encouragements and advices from Nijmegen to Seattle. I need to thank many other DIAGies including old and new colleagues: Mark van Grinsven, Freerk Venhuizen, Eva van Rikxoort, Carl Shneider, Paul Konstantin Gerke, Wendy van de Ven, Alberto Traverso, Zijian Bian, Wendelien Sanderink, Bart Liefers, Sven Lafebre, Sil van de Leemput, Anton Schreuder. I'll never forget the many wonderful lunches and fun activities we have done together.

Our team at the digital pathology group had a rapid growth. Oscar and Maschenka were among the first colleagues to join us. Oscar thank you so much for being my paranymph. You are a great friend and we share so many wonderful memories together (Prague, Heidelburg, Australia, etc.). The trip to Prague was a major eye-opening event for us. Apart from the CAMELYON16 challenge, we learned a lot about Khayam, the master poet who believed life is about wine and women (I still get a laugh when I reminisce about the afternoon we spent with Parham in Pragueto learn all about these). I am so happy for you to become a father of an adorable, cute, and sweet son. Oscar, I wish you the best of luck with your PhD studies. Maschenka, as soon as you joined our group you turned into one of the main contributors in my works. I learned a lot about different aspects of breast pathology from you. The addition of new brilliant members made our group stronger. Francesco, David, Peter, and Wouter, I wish you the best of luck with your work. We had many nice memories together. I want to thank Guido Zuitdhof who did his master thesis under my supervision. Guido was one of the most outstanding master students I encountered at Radboud. Guido had enormous contribution to the work in chapter 6. I need to thank the other members and collaborators of our group, Marcory, Dan (the picture of kitties is reserved for you), Thomas, and John-Melle.

Many thanks to Roland Holland, who is an inexhaustible fount of knowledge in breast pathology and radiology. Roland graciously contributed his time and insights about my work in chapters 5 and 8. I fondly remember him getting super excited and encouraging our work on identifying tumor-associated stroma (chapter 8).

During my PhD, I had a research visit to BeckLab at Harvard for 6 months. It was a wonderful learning experience for me and the research outcome had a lot of impact on this thesis. I would like to start with thanking Andrew Beck who made this

visit possible. Andy, I appreciate all your support, assistance, and encouragement. I am very happy to see that our joint work resulted in two journal paper submissions and one conference paper. During my visit at BeckLab, I have been very fortunate to work with many remarkably inspiring people at Harvard Medical School. I would like to thank Dayong Wang, Francisco Beca, and Korsuk Srirnukunwata for the thoughtful discussions and valuable comments on my work. I especially would like to thank Mitko Veta who helped me with his expertise, friendship, and encouragement. Mitko had a major contribution in chapter 7 of my thesis. Mitko, you are a wonderful friend and your presence at HMS made my experience more remarkable. Very importantly, I dig your Macedonian final touch on cooking pasta. I also would like to thank Jan, Ben, Octavian, Humayun, Jong Cheol, Manan, Jimmy, and Aditya who made my experience at Harvard a very pleasant one.

I would like to thank my close collaborators at the National Cancer Institute, National Institutes of Health (NIH). It was a great pleasure to collaborate with Maeve mullooly, Gretchen Gierach, Ruth pfeiffer, and Maya palakal at the NIH. It was also a distinct pleasure for me to work with Mark Sherman from Mayo Clinic. Thank you Mark for all the insightful discussion and feedback on my work and the enormous contributions to the content of Chapter 8. I found our work on analysis of Tumor-associated stroma one of the most interesting topics of my PhD with potentially the highest impact on patients' health and I am very happy for our fruitful collaboration.

I need to thank the wonderful secretaries of DIAG and Radiology department, in particular, Charlotte Neger, Solange Estourgie, and Leonie Vos. You have always been incredibly helpful and I appreciate the time you took to assist me.

I would also like to extend thanks to some people from the other groups at the radiology department, who I had the pleasure of interacting with and learning from, including but not limited to Nassim, Houshang, Tom Peeters, Marnix, Isabel, Martin van Amerongen, Frits, and Mirriam, many of whom have been my office-mates as well. Nassim, thank you so much for all your friendship, supports, encouragements, and all the cheerful and memorable moments. Together with Kasra, Houshang, Mohsen, and Masoumeh we have had great times together. I wish you all a lifetime of happiness and success.

And those are just the people who directly influenced my research. There is no way I can ever repay the amazing people in my life who got me this far. I am grateful for my closest friends living all around the world who have been supportive of me in all conditions. Mohammad in Germany, Mehdi and Sina in Canada, Ramtin and Irene in Sweden, and Ramin in Denmark. A big thank you to my encouraging and supportive brother, Bahador. And to my parents for their unconditional love, nurturing, patience, and continued and faithful support throughout my life.

Curriculum Vitae

Babak Ehteshami Bejnordi was born in Rasht, Iran, on September 18th, 1986. He studied Electronic Engineering as his undergraduate major at the University of Guilan in Iran. In 2013, he received his Master's degree in Electrical Engineering (specializing in Biomedical Engineering) from Chalmers University of Technology in Sweden. His Master thesis project was part of an automated cytology project involving collaboration with Uppsala University. The thesis was entitled "Chromatin pattern analysis of cell nuclei for improved cervical cancer screening" and concerned the application of machine learning and image analysis techniques to Pap-stained cervical smear images for the detection of malignancy-associated changes. In April 2013, he started his PhD project at the Diagnostic Image Analysis Group under the supervision of Prof. dr. Nico Karssemeijer, Dr. Jeroen van der Laak, and Dr. Geert Litjens. His thesis focused on the development of machine learning and image analysis techniques for computerized diagnosis of breast cancer in histopathological images. It constituted part of two larger research initiatives known as VPH-PRISM project, Virtual Physiological Human: Personalized Predictive Breast Cancer Therapy Through Integrated Tissue Micro-Structure Modeling, and the cross-sectional Breast Radiology Evaluation and Study of Tissues (BREAST) Stamp Project. During his Ph.D., between June to November 2016, he joined BeckLab at Harvard Medical School as a visiting researcher where he worked under supervision of Dr. Andrew Beck and applied deep neural networks for analysis of breast cancer stroma in pathology images. The results of his Ph.D. research are presented in this thesis.